DOCUMENT RESUME

ED 276 737 TM 860 625

AUTHOR Kingston, Neal M.; Holland, Paul W.

TITLE Alternative Methods of Equating the GRE General Test.

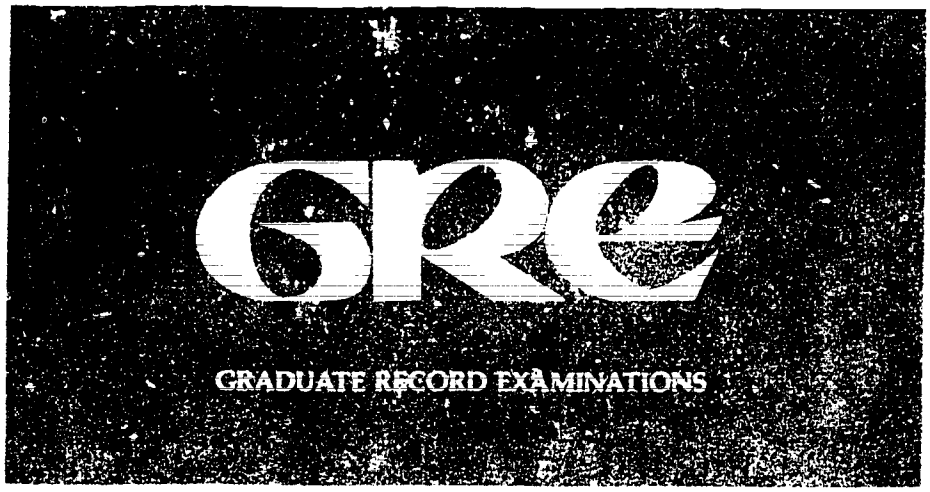
INSTITUTION Educational Testing Service, Princeton, N.J.

SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-86-16; GREB-PR-81-16P
PUB DATE May 86
NOTE 58p.; Appendices contain small print.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Data Collection; *Equated Scores; Error of Measurement; Estimation (Mathematics); Higher Education; *Latent Trait Theory; *Mathematical Models; Mathematics Tests; Research Design; Scoring; *Statistical Bias; Statistical Studies; Test Construction; Verbal Tests
IDENTIFIERS Analytical Tests; *Graduate Record Examinations

ABSTRACT

This study addresses the test-disclosure-related need for more Graduate Record Examinations (GRE) General Test editions in a situation where the number of examinees is stable or declining. Equating is used to guarantee that examinees of different test editions are treated equitably. The data collection designs used in this study were: (1) Nonrandom Group External Anchor Test (NREAT); and (2) Random Group, Preoperational Section (RPOS). Bias and root mean squared error were calculated for the verbal, quantitative, and analytical GRE measures. Item response theory (IRT) and linear equating definitions were applied. In using RPOS or IRT, a high bias and root mean squared error were shown for equating the verbal and analytical measures, whereas a small amount of bias and a moderate amount of root mean squared error were shown for equating the quantitative measure. In using NREAT, quantitative equatings had moderate amounts of both bias and root mean squared error. Small amounts of bias and root mean squared error were shown in equating for the verbal measure. A list of references, data tables and figures, linear equating models, and notes on other equatings are appended. (JAZ)



ALTERNATIVE METHODS OF EQUATING THE GRE GENERAL TEST

Neal M. Kingston
Paul W. Holland

GRE Board Professional Report GREB No. 81-16P
ETS Research Report 86-16

May 1986

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Alternative Methods of Equating the GRE General Test

Neal M. Kingston
Paul W. Holland

GRE Board Professional Report No. 81-16P

May 1986

Acknowledgements

Financial support from the Graduate Record Examinations Board and Educational Testing Service is gratefully acknowledged. Our thanks to our many colleagues who made this research possible. The original plan for this research was designed by E. Elizabeth Stewart with the assistance of Madeline Wallmark and several other consultants. Many of the analyses were supervised or performed by Madeline Wallmark, Dorothy Thayer, and Craig Mills. We are especially grateful for the organizational and programming assistance of Louann Benton. We also thank Frederic Lord, Martha Stocking, and Marilyn Wingersky with whom we consulted many times and several colleagues who reviewed an earlier draft of this paper. We wish to thank especially E. Elizabeth Stewart for her insightful comments. Nonetheless, the opinions expressed herein are solely those of the authors.

ABSTRACT

The original purpose of this study was to address the test-disclosure-related need to introduce more Graduate Record Examinations (GRE) General Test editions each year than formerly, in a context of stable, or possibly declining examinee volume. The legislative conditions that created this initial concern regarding test equating have abated. However, several of the test equating models considered in this research might provide other advantages to the GRE Program. These potential advantages are listed in the body of the report.

Equating can be considered to consist of three parts: (1) a data collection design, (2) an operational definition of the equating transformation, and (3) the specific statistical estimation techniques used to obtain the equating transformation. Currently, the GRE General Test collects data using an equivalent groups design. Typically, a linear equating method is used, and the specific estimation technique is setting means and standard deviations equal.

For this research, two other data collection designs were studied: nonrandom group, external anchor test, and random group, preoperational section. Both item response theory (IRT) and linear equating definitions were used. IRT true score equating was based on item statistics for the three-parameter logistic model as estimated using LOGIST. Linear models included section pre-equating using the EM algorithm, Tucker's observed score model, and several true score models developed by Tucker and Levine. For each of the three GRE measures, verbal, quantitative and analytical, all equating methods were assessed for bias and root mean squared error by equating a test edition to itself through a chain with six equating links.

Bias and root mean squared error were extremely large for equating the verbal and analytical measures using section pre-equating or IRT equating with data based on the random group preoperational section data collection design. For the quantitative measure, this data collection design produced a small amount of bias, but moderate amount of root mean squared error.

Using the nonrandom group, external anchor test data collection design, quantitative equatings had moderate amounts of both bias and root mean squared error. Verbal nonrandom group, external anchor test equatings showed relatively small amounts of bias and root mean squared error, with the Tucker observed score model performing particularly well. Bias was small for the analytical anchor test equatings, and root mean squared error ranged from small to moderate.

All nonrandom group, external anchor test methods worked about as well in practice for the verbal measure as the currently used random group method does in theory. The current random group method, however, has never been subjected to an empirical check comparable to that used in this study for the experimental equating methods. Two anchor test

methods, Tucker 2 True and Levine, appear to have worked as well in practice for the analytical measure as the random group method does in theory.

A possible explanation for the generally poor results for the random group, preoperational section data collection design based equatings was the constant use of the last section of the test to collect equating data. It may be, now that the sections of the GRE General Test are administered in various orders in different editions of the test, that the extreme bias found in this study for the verbal and analytical random group preoperational section equatings will disappear or at least be substantially diminished.

INTRODUCTION

Purpose of this study

The original purpose of this study was to address the test-disclosure-related need to introduce more Graduate Record Examinations (GRE) General Test editions each year than formerly, in a context of stable, or possibly declining examinee volume. Since then, the legislative conditions that created the initial concern regarding test equating have abated. However, several of the test equating models considered in this research might provide other advantages to the GRE Program, such as: improved test security, greater accuracy of equating, shorter time-schedule requirements for score reporting, additional test analysis information, and possible improvement of the test development process.

Equating

Test developers usually try to make the various editions of a test interchangeable with regard to content coverage, item format, and difficulty so that examinees are neither advantaged nor disadvantaged by the edition of the test they happen to take. Unfortunately, because of the large number of constraints under which test developers operate and because of the quality of the statistical information that is available at the time editions of a test are constructed, inevitably, some test editions are easier than others. To make sure that groups of examinees taking different editions of a test are treated equitably, statistical techniques known as test equating are used to adjust scores on each edition of the test so that they are comparable to scores earned on other editions of the test.

There are several different equating models used by psychometricians. These models make different assumptions about the data they use and vary in their appropriateness for any given examination, such as the verbal, quantitative, and analytical measures of the GRE General Test. There are three major aspects of any equating model: (1) the data collection design, (2) the operational definition of the equating transformation, and (3) the specific statistical estimation techniques used to obtain the equating transformation.

Data collection designs. Two data collection designs were used specifically for this study. For the first design, referred to as NonRandom group External Anchor Test (NREAT), two editions of a test were administered, one to each of two nonrandom groups consisting of examinees who chose to take the test on one or the other of two test administration dates, and a common short test that did not count toward the examinees' scores was administered to both groups. This design is sometimes referred to as Design IV from Angoff (1984).

For the second design, referred to as Random Group, Pre-Operational Section (RPOS), essentially, one test was administered to a group of examinees. That group was further divided into two equivalent subgroups through the spiraling of test booklets. That is, different versions of that test edition were packaged in an alternating fashion (e.g., 1,2; 1,2; 1,2; ...). Research has shown that spiraling results in

essentially equivalent groups, sometimes even more effectively (because of a stratification effect) than does true random assignment. In addition, one subgroup received one-half of a second edition of the test, and the other subgroup received the other half of the second edition. The two half-tests were designed to be as similar as possible in content and difficulty. For further information on this design, see Holland and Thayer (1981, 1985), Holland and Wightman (1982), or Petersen, Hoover, and Kolen (in press).

The data collection design used currently for operational GRE General Test equatings is referred to as Random Groups (RG). Two editions of the test are given, one to each of two random, or otherwise equivalent, groups. The GRE Program regularly uses spiraling to ensure equivalence of the two (or more) groups. This design was used in this research to provide some comparison equatings.

More detail regarding the data collection designs used in this study is given in the Procedures section of this report.

Equating transformations. Three operational definitions of the equating transformation are commonly used: (1) linear equating, which provides a transformation such that scores from two tests will be considered equated if they correspond to the same number of standard deviations from the mean in some population of examinees, (2) equipercentile equating, which provides a transformation such that scores from two tests will be considered equated if they correspond to the same percentile rank in a specified population of examinees, and (3) item response theory (IRT) equating, which provides a transformation such that scores from two tests will be considered equated if they correspond to the same level of the latent trait underlying the two tests. Only linear and IRT transformations were used in this study.

Statistical estimation techniques. A number of different techniques have been developed to estimate the intercept and slope parameters for a linear equating. Each technique attempts to estimate the first two moments of the score distributions for an old edition (one whose scores are already on scale) and a new edition of the test on some common group of examinees. These estimation techniques differ in the assumptions that they require. A primary difference is that some linear methods estimate the means and standard deviations of observed scores and others estimate the true score moments. Estimating true score moments is considered particularly appropriate when the two editions of the test to be equated have been administered to groups with very different ability distributions (Angoff, 1984, p. 113). In this study, various statistical estimation techniques, which will be described later in this report, were used with the NREAT data collection design, and one, EM algorithm, to estimate the first two observed score moments (Holland and Wightman, 1982) was used with the RPOS data collection design. This latter method is commonly referred to as section pre-equating, or SPE.

This study investigated only one IRT equating method, IRT true score equating (Lord, 1980, pp. 199-200). There are three aspects of

statistical estimation for IRT true score equating: (1) estimating item and person parameters, (2) putting the parameter estimates from separate calibration runs on a single scale, and (3) setting equal the true scores that correspond to the same level of the latent ability, theta.

The GRE General Test[1]

The GRE General Test measures and yields separate scores for the general verbal, quantitative, and analytical abilities students should have acquired to be successful at the graduate level of education. Scores for each measure are based on the number of correct answers and are scaled to fall between 200 and 800. The test consists of seven 30-minute sections of multiple-choice questions. At the time data were collected for this research, sections 1 and 2 constituted the verbal measure; sections 3 and 4, the quantitative measure; and sections 5 and 6, the analytical measure. The remaining section does not count toward any of the reported scores, and usually consists of verbal, quantitative, or analytical pretest items. For this research, the remaining section was used to collect data for the equating experiments. The specially constructed versions of this section are described in the Research Design section of this report. Since the data for this research were collected, the ordering of sections of the General Test has changed. For current editions of the GRE General Test, the seven sections may be arranged in various orders.

The verbal measure employs four types of questions: antonyms, analogies, sentence completions, and reading comprehension sets. The quantitative measure employs three type of questions: discrete quantitative questions, data interpretation questions, and quantitative comparison questions. The quantitative questions measure basic mathematical skills, understanding of elementary mathematical concepts, and the ability to reason quantitatively and solve problems in a quantitative setting. These questions require arithmetic, algebra, and geometry at a level not beyond that taught in a first high school level course.

The analytical measure employs two types of questions: analytical reasoning and logical reasoning. Analytical reasoning questions test the ability to understand a given structure of arbitrary relationships among fictitious persons, places, things, or events; to deduce new information from the relationships given; and to assess the conditions used to establish the structure of relationships. Logical reasoning questions test the ability to understand, analyze, and evaluate arguments: recognizing the assumptions on which an argument is based, drawing conclusions from given premises, inferring material missing from given passages, applying to one argument principles governing another, identifying methods of argument, evaluating arguments and counter-arguments, and analyzing evidence.

---

[1] This section of the paper was adapted from the Guide to the Use of the Graduate Record Examinations Program 1985-86 (ETS, 1985a).

Additional information on the content of the GRE General Test and examples of the various item types can be found in the GRE Bulletin (ETS, 1985b).

RESEARCH DESIGN

In this section the database (test editions and examinee samples) is described, and the various procedures used in this research are detailed. The equating models used and the assumptions upon which they are based are explained. The rationale for the criterion used to judge the adequacy of the equating models is developed.

Database

Six editions of the GRE General Test were administered on seven different occasions; the edition given at the first and last administration was the same.

Test editions. For ease of reading, the six editions of the GRE administered as part of this research will be referred to in this report as E1, E2, E3, E4, E5, and E6. (The ETS designations for these test editions are 3DGR3, 3DGR1, 3DGR2, 3EGR1, 3EGR4, and 3EGR2, respectively, for the verbal and quantitative measures. For the analytical measure, edition 3EGR3 was used instead of 3EGR4.) One of several different experimental sections was administered along with each edition. These experimental sections were used as either anchor tests for the NREAT equating data collection design or as pre-operational sections for the RPOS data collection design. The use of these data for equating will be further explained later in this section of the report.

Table 1 describes the characteristics of the verbal, quantitative, and analytical measures for each test edition. It shows when each edition was administered as part of this research. For each measure it gives the number of items contributing to the reported score for that edition and the mean and standard deviation of the difficulty[1] of the items in that measure. In addition, the number of items in each experimental section and their difficulty are presented. Appendix A presents (among other information) the ETS form and subform designations and codes for each test.

Insert Table 1 About Here

Examinee samples. Samples consisted of all examinees who took the appropriate test editions at one of the seven administrations, with the exception of all of the following:

- examinees who had taken the GRE General Test more than once and who received any of the same test sections at two or more administrations,

---

[1] Item difficulty is presented in terms of equated deltas, that is deltas put onto a common scale for a given measure for all test editions. See Henrysson (1971, pp. 139-140) for a description of the delta statistic.

- examinees who did not respond to at least five items in each of the seven sections of the test,
- examinees without an item response record (i.e., any examinees whose answer sheets were not machine scored), and
- examinees who took the test at a center for which an administrative irregularity was reported.

Tables 2, 3, and 4 present information regarding the examinees tested as part of this research -- the sample sizes and means and standard deviations of their scaled verbal, quantitative, or analytical scores for the subgroups tested at each administration. The statistics are based on only those examinees used in each equating. Appendix A presents (among other things) the number of examinees in the sample for each external anchor test and each preoperational section. Anchor test samples ranged from 3,583 to 4,408 examinees. Preoperational section samples ranged from 1,745 to 2,561. The approximately two-to-one ratio of sample sizes for anchor test and preoperational section samples was planned, since two preoperational section samples, but only one anchor test sample, are needed for each score being equated at a given test administration. Note, however, that all data to be used for equating based on preoperational sections can be collected in one administration, but data for equating using external anchor tests must be gathered from two test administrations.

Insert Tables 2, 3, and 4 About Here

Figures 1 and 2 present the NREAT and RPOS data collection designs, respectively. For the NREAT design, at each administration (other than the first and last) six forms of the edition administered were spiraled, two for each of the General Test measures. In each pair, the operational test edition was the same, but one form contained in the seventh section an anchor test (containing items of the same types as the measure being equated) in common with the previously administered test edition, and the other contained an anchor test in common with a test edition scheduled for a future date.

For the RPOS data collection design, at each administration six forms of the test were spiraled, two for each of the three General Test measures. In each pair the operational test edition was the same, but one form contained in the seventh section one of the two operational sections of either the verbal, quantitative, or analytical measure from a previously administered edition, and the other form contained the other operational section. Note, this is the opposite of what is normally done for an RPOS data collection design. Usually, these sections would contain halves of future editions instead of previous editions.

Insert Figures 1 and 2 About Here

Procedures

Equating methods. Results from seven different equating methods are presented in this report -- one RG method: setting means and standard deviations equal; four NREAT methods: Tucker, Tucker True 2, Levine (equally reliable, or unequally reliable, as appropriate), and IRT Anchor Test True Score with theta metric set using concurrent calibration; and two RPOS methods: IRT Preoperational True Score with theta metric set using concurrent calibration, and EM algorithm to estimate means and standard deviations. Three-parameter logistic IRT estimates were performed with the program LOGIST (Wingersky, Barton, & Lord, 1982). A brief overview of these methods is given in Table 5, and a detailed description of the linear methods is presented in Appendix B (Appendix B is adpated from appendix A of Marco, Petersen, & Stewart, 1983). Detailed information on IRT true score equating is available in Lord (1980, chapter 13) and Hambleton and Swaminathan (1985, chapter 10). Information on the use of the EM algorithm for equating with RPOS data collection designs is available in Holland and Wightman (1982).

Insert Table 5 About Here

Some additional linear equating methods were used in early stages of this study as was a second method of establishing a common IRT metric. Appendix C presents some notes on these methods.

Assessing the adequacy of equating methods. A good equating method should have certain characteristics. As with many other statistical estimation techniques, these desirable characteristics include minimal bias and mean squared error. Assessing the bias and mean squared error of one or more equating methods, however, first requires one to know the true equating relationship between test editions. In most real-life equating situations, this is not possible. For the purpose of this research, such a criterion was constructed.

For each equating method in this study, a chain of six equatings was performed. E2 was equated to E1, E3 was equated to E2, E4 to E3, E5 to E4, E6 to E5, and then finally E1 was administered again and equated to E6. If the function that equates E2 to E1 is called $f(x)$, and the function that equates E3 to E2 is $g(x)$, then the composite function $g(f(x))$ will put scores from E3 on the E1 scale. Likewise, if $h(x)$ equates E4 to E3, $i(x)$ E5 to E4, etc., then $k(j(i(h(g(f(x))))))$ equates E1 to E1 through the chain (or circle) of six equatings. Since the equating of E1 to E1 should be an identity function, it can be determined how close each equating method came to the true equating relationship.

The equating criterion was used in several ways. First, equatings were compared graphically. These graphs were summarized statistically. In doing so, it was decided that inaccuracies in equating were inconsequential if few or no examinees were affected by them, and so discrepancies were weighted at each raw score by the number of examinees (whose data for edition E1 was used in this research) who obtained that score. Two summary statistics were calculated: bias (equation 1), the weighted mean difference between an experimental equating and the criterion equating, and root mean squared error (see equation 2), equivalent to the weighted mean standard error of equating. Note that root mean squared error includes bias and thus can never be smaller than bias.

$$\text{bias} = \left(\left[\sum_{i=1}^{n} X_{2i} f_i\right] / \sum_{i=1}^{n} f_i\right) - \left[\sum_{i=1}^{n} X_{1i} f_i\right] / \sum_{i=1}^{n} f_i \qquad (1)$$

$$= (\bar{X}_{2i}) - \bar{X}_{1i}$$

$$= \bar{d}$$

$$\text{RMSE} = \sqrt{\left(\sum_{i=1}^{n} [X_{2i} - X_{1i}] f_i\right) / \sum_{i=1}^{n} f_i} \qquad (2)$$

$$= \sqrt{\left(\sum_{i=1}^{n} d_i f_i\right) / \sum_{i=1}^{n} f_i}$$

where n is the maximum obtained score for the measure,

X_{2i} is the score (equated through the six-link chain) corresponding to raw score i for the April 1983 administration of form E1,

f_i is the frequency of raw scores i in the October 1981 group,

X_{1i} is score corresponding to raw score i for the October 1981 administration of form E1.

In addition to the calculation of bias and root mean squared error on the raw score metric, to facilitate comparisons of equating methods across the three General Test measures, bias and root mean squared error (RMSE) were standardized by dividing each by the scaled score standard deviation for the appropriate test score. Also, to provide a context familiar to most score users, bias and root mean squared error were transformed to the appropriate GRE scaled score metric (verbal, quantitative, or analytical) by multiplying them by the slope of the scaling function used to place raw scores for edition E1 on the GRE score

scale. For the verbal measure this transformation is nonlinear, so a linear approximation was used. This approximation differed from the actual scaling primarily at the high end of the scale where there are few data.

RESULTS

Figures 3, 4, and 5 graphically present the equating lines (raw score new to raw score old) for the six methods used in this study, for the verbal, quantitative, and analytical measures, respectively. Figures 6, 7, and 8 present the raw score differences between each equating line and the true equating function. So, for example, for the verbal measure, the SPE method would convert a raw score of 72 to a 66 (even though the equating should have been an identity function yielding an equated score of 72). Sixty-six minus 72 is negative six, and this can be seen in Figure 6. Essentially, these difference graphs simply magnify the discrepancies between equating methods. Note that the scale for Figure 6 is different from that for Figures 7 and 8.

Insert Figures 3 through 8 About Here

Figures 3 through 8 show that for the verbal and analytical equatings, the two methods based on the RPOS data collection design worked least well. For the quantitative equatings, the graphs do not show a readily discernible difference in the quality of the equating methods.

Table 6 presents the bias and root mean squared error for the six equating methods for each General Test measure. Table 7 presents bias and root mean squared error in the GRE General Test scaled score metric, so that equating error can be viewed in a context familiar to GRE score users. Table 8 presents the standardized bias and root mean squared error for the six equating methods.

Insert Tables 6, 7 and 8 About Here

Several findings stand out in these tables.

- For the verbal and analytical measures, the equating methods that used a RPOS data collection design (IRT and SPE) had relatively large bias and root mean squared error compared to those that used a nonrandom group external anchor test design.
- For the quantitative measure, however, the absolute standardized bias was least for the SPE and IRT RPOS data collection designs.
- The standardized root mean squared errors for the quantitative equatings were not too different (ranging from .07 to .11) regardless of the data collection design or equating transformation method.

- When making comparisons within each of the three measures that were equated, all models using NREAT equating performed about equally well for the verbal and quantitative measures. For the analytical measure IRT performed somewhat less well than Tucker True 2 and Levine (.10 for IRT, compared to .04 for the other two methods).

- Overall, NREAT equating models did less well in terms of both bias and root mean squared error for the quantitative equatings than for the verbal or analytical equatings.

- For the verbal equatings, for each equating method, the root mean squared error was substantially accounted for by a consistent negative bias: that is, items in the preoperational section were systematically more difficult than when they appeared in an earlier operational section. This bias was small, however, for all of the NREAT equating methods.

- For the quantitative equatings based on NREAT methods, the root mean squared error was primarily accounted for by a consistent positive bias. For the RPOS data collection design, however, the bias was small and inconsistent, and did not account for substantial amounts of the root mean squared error.

- Overall, the analytical equatings based on NREAT methods, bias was small and accounted for very little of the root mean squared error. For the RPOS methods, a consistent large negative bias accounted for most of the root mean squared error.

DISCUSSION

Comparison of empirical root mean squared error with the standard error of a chain of operational equatings

In order to answer the question, "did any of the NREAT or RPOS equatings work well enough," some context is needed. One such context is the standard error of a chain of equatings based on the method currently used to equate the GRE General Test: random groups, setting means and standard deviations equal (Angoff, 1984, design IA-1, pp. 94-97). The standard error of equating is affected by the size of the sample on which the equating experiment is performed. For operational GRE General Test equatings, the samples for each edition of the test range from 10,000 to 20,000 and thus the total sample size ranges from 20,000 to 40,000.

Verbal equatings. For the GRE verbal measure, NREAT equating methods worked quite well, and RPOS methods did not. The scaled score root mean squared error of the chain of equatings for the four NREAT equating methods was about 5 scaled score points. This figure can be compared to the standard error of a chain of six equatings for the operational equating method -- random groups, setting means and standard deviations equal. The operational standard error was estimated using Lord's formula for the standard error of a single RG means and standard deviations equating (Lord, 1950; Angoff, 1984, p. 97) and Theorem 6, by Braun and Holland (1982), for the standard error of a chain of equatings. Assuming that the slopes of the raw score to raw score equating functions are close to one, the standard deviation of the equated scores for each new form group was 123 (the average for the equating groups used in this research), the total number of examinees upon which each operational equating is based is 30,000, and test scores are normally distributed, then the median standard error of equating for the chain of six equatings is about 4 scaled score points. (The median standard error is the standard error at .675 standard deviations from the mean. Under the previously mentioned assumptions, this median is comparable to the empirical weighted root mean squared error.) If the total number of examinees upon which each equating was based were 20,000 or 40,000, the median standard error of the chain of equatings would be about 5 or 3 scaled score points, respectively. Of course, the figures for the median standard error of equating depend on the previously listed assumptions as well as on the assumptions of RG linear equating. In practice these assumptions may be violated. This would probably increase the empirical RG linear standard errors of equating to a value at least somewhat larger than the theoretically derived numbers presented here. Thus, for the purpose of evaluating the root mean squared error, the median standard errors should be considered conservatively low.

For the two RPOS equating methods, the average verbal equating root mean squared error was 26 scaled score points, considerably worse than the estimated standard error of the chain of operational equatings.

Quantitative equatings. For the quantitative measure, all equating methods had about the same root mean squared error, on the average about 13 scaled score points. Assuming a scaled score standard deviation of

133, a normal distribution, and sample sizes of 20,000, 30,000, and 40,000, the median standard errors of equating for the chain of six quantitative equatings using the operational method are about 5, 4, and 4 scaled score points, respectively. Thus, the NREAT equatings do not appear to have worked as well for the quantitative measure as they did for the verbal measure. And, although for the quantitative measure the RPOS equating methods worked as well as did the NREAT equatings with regard to root mean squared error and worked better with regard to bias, they did not perform as well as one would expect a random groups linear equating to perform.

Analytical equatings. For the analytical measure the IRT NREAT equating had the largest root mean squared error, 13 scaled score points. The linear NREAT equating methods had root mean squared errors ranging from 5 to 9 scaled score points. The two RPOS equatings had an average root mean squared error of 26 points. With a scaled score standard deviation of 126 and sample sizes of 20,000, 30,000, and 40,000, the median standard errors of the chain of analytical RG linear equatings would be 5, 4, and 4 scaled score points, respectively. Therefore, the linear NREAT methods did reasonably well, but the IRT NREAT equatings, and to a much larger extent the RPOS equatings, did poorly.

Factors That May Have Affected These Results

Only for the verbal measure NREAT equatings did the root mean squared error appear reasonable in light of the theoretical standard error of the current GRE equating procedure. This might be due to any of at least three factors. First, the samples used in this study are considerably smaller than the samples used in the operational equating of GRE scores.[1] This might be compensated for, however, by the increased power of NREAT and RPOS equating designs. Second, sampling error might have produced large root mean squared errors in the groups used in this study even though in another set of equatings the root mean squared errors might be smaller. Third, the root mean squared errors of the experimental equatings are based on real data and not on statistical assumptions. The effect of real data is likely to be an increase in the size of the root mean squared error of the operational procedure beyond the theoretical standard error of equating. If the empirical root mean squared error of a chain of random group, means and standard deviations equatings were calculated, it might also be somewhat larger than the theoretical standard error of equating. This might occur because of violations of the assumptions of the equating model. In particular, to some extent examinees are advantaged if they have previously taken the same edition of a test (Kingston & Turner, 1984). This can occur for the old edition, but not for the new edition, in an RG equating.

[1] The choice of sample sizes in this study was intentional and reflects administrative constraints such as current administration volumes and pretesting needs

It should be noted that if NREAT or RPOS data collection designs were used operationally for the GRE General Test, it is likely that double-part score equating would be used. For NREAT data collection, this would entail using two anchor tests in order to equate to two different old editions of the test and then averaging the two equatings. Likewise, for an RPOS data collection design, a new edition of the test would be preoperationally equated to two different old editions, and the average of the two equatings would be used. Although the statistical properties of double-part score equating are not well understood, such an equating would be expected to have reduced root mean squared error and would be expected to reduce certain sources of bias (although not the sources of bias that appear to have affected the RPOS equatings in this study).

Effect of smaller sample sizes. If the standard errors were calculated for the equatings performed in this study, then the effect of the smaller samples used in this research could be addressed directly. Unfortunately, no method has yet been devised to assess the standard error for IRT or section pre-equating. Several methods have been proposed for estimating the standard error of NREAT linear equatings (Lord, 1975; Kolen, 1985). Table 9 presents for the verbal, quantitative, and analytical measures, the standard error of the chain of Tucker equatings based on the delta method developed by Lord. The median standard errors (assuming normality of the score distributions) are 4, 4, and 5 scaled score points, respectively, for the verbal, quantitative, and analytical measures. Thus, at least for the linear NREAT equatings, the smaller sample sizes appear to have been compensated for by the reduction of sampling error from the use of anchor test data. That is, the median standard errors of the linear NREAT equatings based on samples of about 4,000 are about the same as the median standard errors of the random group, means and standard deviations equatings based on samples of 20,000 to 30,000.

Insert Table 9 about Here

The RPOS equatings were based on samples approximately one-half the size of the NREAT samples. Could this explain the particularly poor performance, at least in the case of the verbal and quantitative measures, of the RPOS equating methods? Since no standard errors could be estimated for the RPOS equatings, this was assessed by dividing each of the NREAT samples in half, performing Tucker and Levine equatings on each half-sample, and estimating the bias and root mean squared error for each chain of half-sample equatings. Although three of the four half-sample equating chains had greater bias and root mean squared error than did their respective full-sample chains (unexpectedly, one of the Levine half-sample equatings had smaller bias and root mean squared error), all four had considerably smaller bias and root mean squared error than the RPOS equating chains.

In summary, it appears that the performance of the NREAT and RPOS equating methods relative to the standard error of the operational RG equatings cannot be explained by the sample sizes used in this study.

Effect of sampling error. Six different sets of data were used for the equatings in this study. That is, different examinees made up the samples for the NREAT and RPOS groups for each of the three GRE General Test measures. Thus, the results of the 18 equating chains presented in Tables 6 through 8 are based on only six independent sets of data. Still, of those 18 chains, only one (Tucker for the verbal measure) has a root mean squared error smaller than the theoretical standard error of a comparable chain of RG equatings based on total samples of 30,000 examinees. It appears highly unlikely that chance in the selection of samples explains these results.

Effect of real data. All statistical models, including equating models, are based on assumptions that are not strictly met by the data. Thus, standard errors of equating, which are based on these unmet assumptions, are usually unrealistically small compared to corresponding root mean squared errors that are empirically derived. The magnitude of the discrepancy between the standard error of equating and the empirically derived root mean squared error will depend on the magnitude of the discrepancy between the assumptions and the data. The assumptions of the various equating models used in this research are given in Appendix B. Some of these assumptions are untestable, given the available data: for example, the assumption that the regression of total test on equating test for test X in population Q (the population that took test Y) is the same as the regression of test X on the equating test for population P. For other assumptions, such as the local independence assumption of IRT, good methods of testing the assumptions did not exist at the time this research was carried out. Previous research has demonstrated the reasonableness of the three-parameter logistic model for the GRE verbal and quantitative measures (Kingston and Dorans, 1982a). Analysis of item-ability regressions (for an example of such an analysis see Kingston and Dorans, 1985) and a slightly modified Yen's Q_1 statistic (see Yen, 1981, 1984) indicated that the three-parameter logistic model is probably reasonable for the current GRE General Test.

One assumption buried in RPOS equating is that examinee responses to items will be the same when the items appear in the preoperational sections and when they appear in the operational test. Examinees' responding behavior might vary, for example, if they knew that the preoperational items did not count toward their score and therefore they decided not to waste too much time and energy on those items. More generally, behavior might vary if there were any kind of context or location effect, perhaps caused by fatigue or practice in one setting, but not in the other (Kingston and Dorans, 1984; Whitely and Dawis, 1976; Yen, 1980).

In this research, preoperational data were always gathered in the seventh (last) section of the test. These "preoperational" data were for test material from a previously administered edition (that is, the old

form in the equating relationship) in which the items always appeared prior to the last section. The considerable negative bias found for the verbal and analytical RPOS equatings indicates that for a given raw score, scaled scores for E1 when administered operationally in April 1983 were lower than for E1 when administered "preoperationally" in December 1981. For two editions of a test, the scaled score for a given raw score will be higher for the more difficult edition. Therefore, the verbal and analytical items appeared more difficult when they were administered in section 7 than when they were administered operationally in sections 1 and 2 (verbal) or 5 and 6 (analytical). Note, however, that this effect, which appears clear from the equatings, is not clear from the mean deltas presented in Table 1.

Some coaching schools have advised examinees to determine which section or sections of a standardized test do not count toward their score and save their energy by not working too hard on those sections (Owen, 1985 pp. 135-136). If a large number of examinees followed this advice, it might explain the results for the verbal and analytical measures. This potential explanation is weakened considerably, however, because it is unclear why this would occur for those measures but not for the quantitative measure. Alternatively, it might be that verbal items (particularly reading comprehension) and analytical items are susceptible to a fatigue effect. If an examinee's attention span diminishes at the end of a long test, this might affect, in particular, items that refer to relatively long passages or that require the juxtaposition of diverse elements of a question. Such an effect would be more likely to influence responses to reading comprehension, analytical reasoning, and logical reasoning items than other item types.

Previous research on the effect of item location on IRT parameter estimates and equating results has been performed on the pre-October 1981 version of the GRE General Test (Kingston and Dorans, 1982b, 1984). That research was based on an RPOS data collection design very similar to the one used in the current study, but it differed in that the test was administered under formula-scoring instructions, the verbal measure was slightly longer and more speeded, and the analytical measure had two additional item types and only a few logical reasoning items. Kingston and Dorans found that location effects were item-type specific. Analysis of explanations and logical diagrams items, two item types no longer used in the analytical measure, showed very large practice effects; that is, they were considerably easier when answered after another section of such items. Reading comprehension items were more difficult for the examinees when they appeared in the preoperational section at the end of the test (the mean difference between b-estimates was .14). Although for two different test editions the magnitude of the effect was consistent, it was statistically significant at the .05 level in only one. Analogy, antonym, and quantitative comparison items all appeared easier in the RPOS position in both editions of the test, but the differences were statistically significant at the .05 level only for antonyms and only in one of the two test editions. Because too few logical reasoning items were administered, no results for that item type are available. Results for the other item types were inconsistent. Given the change in scoring

directions and the inconsistency of these results, the Kingston and Dorans study does not appear to shed too much light on the RPOS equating results.

While Kingston and Dorans did not find consistent statistically significant results for the item types that constitute the verbal and analytical measures of the current General Test, item location effects caused by fatigue or practice cannot be ruled out as an explanation for the RPOS equating results. Even if location effects were too small to be found statistically significant, given the power of the statistical test that would be used, the effects might be consistent, and the sum of such effects over six equatings might be large enough to explain the bias in verbal and analytical RPOS equatings.

Comparison of individual SPE and Tucker equatings. Since for this research the preoperational and operational administration of test items involved different populations, it is not possible to assess directly the magnitude and consistency of any item location effects. One way to assess this indirectly would be to compare each of the six SPE and Tucker equatings in the chain for each measure. This is reasonable for the verbal and analytical measures, since for them the results for the Tucker model were quite good. These results are presented in Table 10. Since for the quantitative measure the Tucker results were not satisfactory, such a comparison will not be presented.

Insert Table 10 about Here

Table 10 shows that for four out of six of the verbal equatings, SPE showed a large negative bias compared to Tucker (items appeared more difficult in their preoperational administration than in their operational administration). For the other two equatings, the bias was positive but smaller in magnitude. SPE shows a negative bias in five of the six analytical equatings. For the E4 to E3 and E6 to E5 equatings the bias is particularly large, 12 and 16 scaled score points, respectively. The one instance of positive bias was also large, about 10 scaled score points.

Comparison of SPE and linear random group equatings. For three of the equating links, E6 to E5, E3 to E2, E2 to E1, there exists for comparison the operational RG linear equating. Such an equating is performed on equivalent samples from a single population and is based on fewer asssumptions than any of the other equating methods presented in this report. Table 11 presents the bias and root mean squared error for the verbal, quantitative, and analytical SPE and Tucker equatings, using the linear random group method as a criterion. It should be noted that this criterion is a relative one. That is, the linear random group equating suffers from some amount of sampling error, it may be more population dependent than some other equating methods, and it assumes linearity of equating relationships (as does SPE).

Insert Table 11 about Here

From Table 11 it appears that for the verbal measure, bias was introduced in the E3 to E2 section pre-equating and not in either of the other two equatings. For the quantitative measure, it seems that a moderate negative bias is introduced in two of the three equatings, but a large positive bias is introduced in the E6 to E5 equating. In fact, for all other linear NREAT and RPOS quantitative equatings (data not presented here), the E6 to E5 equatings had large positive bias (between 8 and 11 scaled score points) and root mean squared error (between 9 and 11 points) compared to the linear RG equating.

Data for the E6 to E5 equatings were studied closely to try to ascertain why they appeared to produce so much bias. The two test editions were matched well on difficulty and had essentially equal reliability. Test analysis data from the original administrations of the two editions confirmed their statistical parallelism (Wallmark, 1982, 1984). The samples for the NREAT equatings were well matched with regard to anchor test score distributions. Although no cause for the poor results for the NREAT equatings is apparent, Table 3 provides evidence as to the possible culprit for the RPOS results. The mean quantitative scores on edition E6 administered in February 1983 were different for the two groups who took the "preoperational" half versions of E5. The means of 526 and 534 differ by about 2.6 standard errors of the mean: a difference that should occur less than one time out of one hundred by chance. The standard deviations for the two groups also differed: 132 versus 129. Although these differences may well be due to chance, this may have affected the quality of the equatings.

For the analytical measure, a very large amount of negative bias was introduced in the E6 to E5 SPE equating (about 15 scaled score points). Also, a moderate amount of bias was introduced in both the SPE and Tucker E2 to E1 equatings. No cause for this bias is apparent. Less bias was introduced in the E3 to E2 equating.

SUMMARY AND CONCLUSIONS

NREAT equating, both IRT based and traditional, worked well for the GRE verbal measure. The average estimated bias for the four equating methods was about 4 scaled score points and the root mean squared error was about 5 points. This appears to be small, given that the verbal measure scaled score standard error of measurement is about 34 points (ETS, 1985a), and that score users are advised not to make distinctions between individuals based on small differences. Also, the NREAT verbal measure root mean squared error is about the same size as the theoretical value for the operational RG General Test equating method. NREAT double-part score equating for the verbal measure might well be psychometrically somewhat superior to the current GRE equating method.

For the quantitative and analytical measures, the NREAT equatings worked somewhat less well. There was essentially no bias over the chain of six equatings for the analytical measure, but the root mean squared error, depending on the equating method, varied between 5 and 13 scaled score points (average root mean squared error was 8 points). Since there is no discernible bias, groups of examinees would not be disadvantaged because of the test edition they happened to take. The random error, however, would be expected to inflate by about 20 percent the alternate form standard error of measurement beyond the internal consistency based estimate. This is somewhat larger than would be expected to occur with the current operational equating method. NREAT double-part score equating is unlikely to be psychometrically superior to the current RG-based equating method for the analytical measure.

The quantitative NREAT equatings had a positive bias of about 12 scaled score points which accounted for almost all of the root mean squared error. This bias is equal to about 30 percent of the 40-point standard error of measurement (compared to about 21 percent for the NREAT verbal bias). The bias and root mean squared error found for the quantitative NREAT equatings is considerably larger than those expected to occur with the operational RG-based equating method, and there is no reason to expect that double-part score NREAT equating would be sufficiently better to justify its use solely on its psychometric merits.

RPOS equating worked poorly for the verbal and analytical sections because of a large negative bias. That is, tests equated from data gathered in the seventh (last) section of the General Test appeared more difficult than when they were operationally administered. There was no such bias, however, for the quantitative RPOS equatings. For the verbal measure the estimated bias was larger for the IRT equatings than for the SPE equatings (27 scaled score points compared to 17 scaled score points), but for the analytical measure the bias was larger for SPE than for IRT (28 points compared to 17 points). These biases are fairly large compared to the standard error of measurement: an average of 72 percent of the standard error of measurement for verbal and an average of 45 percent for analytical. It also appears likely that these biases would continue to propagate and that the verbal and analytical score scales would drift considerably over time.

The quantitative RPOS equatings appeared to work reasonably well. SPE, for the quantitative measure, had a bias of 7 scaled score points, which was small compared to the NREAT biases, as was the root mean squared error of 10 points. The IRT RPOS equatings had no discernible bias, but a root mean squared error of 15 points.

Summary In summary, only the verbal NREAT equatings and none of the RPOS equatings appear to work as well as the operational RG equating works in theory. But all of the NREAT equating methods presented appear to work acceptably well for the analytical measure, and the RPOS equatings appear marginally acceptable for the quantitative measure. Note, however, that the RG methods have not been subjected to an empirical check as have the NREAT and RPOS methods. It is almost certain that RG equating methods will not work quite as well in practice as they do in theory. The important question is whether they work better than the NREAT and RPOS equating methods.

Now that the possible cause of bias in the RPOS equatings has been removed, that is, the constant use of section seven for experimental items, it is likely that RPOS equating methods (based on careful selection of experimental sections in order to balance position) will work better.

Recommendations Two additional research studies are recommended. First, bias and root mean squared error for linear and equipercentile equating using the operational RG data collection design with a six-link equating chain should be designed and performed. The proposed study would provide a meaningful comparison for the results in the current study. Such a study is also recommended by the ETS Standards for Quality and Fairness (ETS, 1983, p. 17), which specifies that testing programs should, "Periodically assess the results of methods used to achieve comparability of scores and evaluate the stability of the score scale."

Item location effects appear to be a thornier problem than previous research has indicated. Additional research leading to models that account for such effects are needed. Such models would increase our knowledge of test-taking behavior and ultimately lead to fairer, more accurate test scores.

REFERENCES

Angoff, W. H. (1984). Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: a mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Educational Testing Service (1983). ETS Standards for quality and fairness. Princeton, NJ: Educational Testing Service.

Educational Testing Service (1985a). Guide to the use of the Graduate Record Examinations Program. Princeton, NJ: Educational Testing Service.

Educational Testing Service (1985b). GRE Information Bulletin. Princeton, NJ: Educational Testing Service.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston: Kluwer-Nijhoff Publishing.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (1981). Section pre-equating the Graduate Record Examination (Program Statistics Research Technical Report No. 81-13). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer D. T. (1985). Section pre-equating in the presence of practice effects. Journal of Educational Statistics, 10, 109-120.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating. A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Kingston, N. M., & Dorans, N. J. (1982a). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. (GRE Board Professional Report 79-12P). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1982b). The effect of the position of an item within a test on item responding behavior: an analysis based on item response theory. (GRE Board Professional Report 79-12bP). Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adoptive testing. Applied Psychological Measurement, 8, 147-154.

Kingston, N.M., & Turner, N.J. (1984). Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test (GREB-83-5P). Princeton, NJ: Education Testing Service.

Kolen, M. J. (1985). Standard errors of Tucker equating. Applied Psychological Measurement, 9, 209-223.

Lord, F. M. (1950). Notes on comparable scales for test scores (RB-50-48). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1975). Automated hypothesis tests and standard errors for non-standard problems. The American Statistician, 29, 56-59.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A large scale evaluation of linear and curvilinear score equating models Volume I (RM-83-2). Princeton, NJ: Educational Testing Service.

Owen, D. (1985). None of the above: Behind the myth of scholastic aptitude. Boston: Houghton Mifflin Company.

Petersen, N. S., Hoover, H. D., & Kolen, M. J. (In press). Scales, norms, and equivalent scores. In R. Linn (Ed.) Educational Measurement (3rd ed.). Washington, DC: American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York, NY: Academic Press.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wallmark, M. (1982). GRE test analysis: Aptitude Test Form 3EGR2. Unpublished Statistical Report (SR-82-83). Princeton, NJ: Educational Testing Service.

Wallmark, M. (1984). GRE test analysis: General test form 3EGR4. Unpublished statistical report (SR-84-05). Princeton, NJ: Educational Testing Service.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two laten trait models. Journal of Educational Measurement, 17, 297-311.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W. M. (1984). Effects of local item dependence on fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Table 1
Description of Test Editions[1]

Test Edition	Admin. Date	Operational # of Items	Operational Mean Delta	Operational S.D. Delta	1st Anchor # of Items	1st Anchor Mean Delta	1st Anchor S.D. Delta	2nd Anchor # of Items	2nd Anchor Mean Delta	2nd Anchor S.D. Delta	1st Preop. # of Items	1st Preop. Mean Delta	1st Preop. S.D. Delta	2nd Preop. # of Items	2nd Preop. Mean Delta	2nd Preop. S.D. Delta
Verbal																
E 1	10/81	72	12.0	2.7	-	-	-	38	12.1	2.6	-	-	-	-	-	-
E 2	12/81	76	11.9	2.4	38	11.9	2.4	38	11.8	2.3	36	11.9	2.5	36	11.8	2.4
E 3	2/82	76	11.8	2.5	38	11.9	2.4	38	11.9	2.5	38	11.8	2.3	38	12.0	2.2
E 4	4/82	76	11.9	2.6	38	12.0	2.5	38	12.3	2.6	38	12.1	2.7	38	11.8	2.2
E 5	10/82	76	11.8	2.9	38	12.1	2.7	38	11.8	2.6	38	11.8	2.8	38	11.9	2.6
E 6	2/83	76	11.9	2.6	38	11.8	2.5	38	11.9	2.5	38	12.1	2.7	38	11.6	2.6
E 1	4/83	72	12.0	2.7	38	12.2	2.3	-	-	-	38	12.2	2.5	38	12.2	2.7
Quantitative																
E 1	10/81	60	11.3	2.7	-	-	-	30	11.1	2.6	-	-	-	-	-	-
E 2	12/81	60	11.3	2.7	30	10.8	2.7	30	11.3	2.6	30	11.1	2.6	30	11.1	2.8
E 3	2/82	60	11.4	2.5	30	11.3	2.4	30	11.2	2.7	30	11.6	2.8	30	11.2	2.6
E 4	4/82	59	11.2	2.5	30	11.2	2.9	30	11.3	2.7	30	11.4	2.4	30	11.5	2.6
E 5	10/82	60	11.3	2.6	30	11.5	2.7	30	11.2	2.5	30	11.4	2.4	29	11.2	2.6
E 6	2/83	59	11.0	2.7	30	11.2	2.5	30	11.4	2.7	30	11.4	2.4	30	11.2	2.5
E 1	4/83	60	11.3	2.7	30	11.3	2.6	-	-	-	30	10.8	2.7	29	11.1	2.7
Analytical																
E 1	10/81	45	13.2	2.1	-	-	-	25	13.4	2.2	-	-	-	-	-	-
E 2	12/81	50	12.4	2.2	25	13.3	2.2	24	12.7	2.0	24	12.7	2.0	21	13.1	1.6
E 3	2/82	50	12.6	2.0	24	12.6	2.2	25	12.4	2.2	25	12.1	2.3	25	12.6	2.1
E 4	4/82	50	12.7	2.3	25	12.6	2.2	25	12.7	2.6	25	12.7	1.9	25	12.8	2.0
E 5	10/82	50	12.5	2.0	25	12.9	2.7	25	12.7	2.1	25	12.5	2.3	25	13.0	2.3
E 6	2/83	50	12.9	2.1	25	12.8	2.2	25	12.7	2.8	25	13.4	1.9	25	12.3	1.9
E 1	4/83	45	13.2	2.1	25	12.8	2.7	-	-	-	25	13.0	2.2	25	12.9	1.6

[1] Anchor - external anchor test
Preop. - preoperational section
Delta - item difficulties on delta scale, equated within GRE measure
Deltas for operational items are based on the item analysis prepared the first time that the test edition was administered; thus the statistics for the two administrations of E1 are the same. Deltas for the anchor test and preoperational sections are based on their administration as part of this research.

Table 2

Description of Samples Used for Verbal Equatings[1]

Test Edition	Admin. Date	1st Anchor	N $\bar{x}$ s	2nd Anchor	N $\bar{x}$ s	1st Preop.	N $\bar{x}$ s	2nd Preop.	N $\bar{x}$ s
E1	10/81	----------		A1	4,408 500 126	----------		----------	
E2	12/81	A1	4,096 473 123	A2	4,180 477 122	E1a	2,062 476 121	E1b	2,076 475 122
E3	2/82	A2	3,746 484 119	A3	3,602 484 118	E2a	1,808 483 120	E2b	1,760 480 120
E4	4/82	A3	3,647 463 124	A4	3,604 462 125	E3a	1,789 464 125	E3b	1,747 465 125
E5	10/82	A4	4,331 518 127	A5	4,230 516 126	E4a	1,713 523 126	E4b	1,654 522 125
E6	2/83	A5	3,825 478 121	A6	3,671 478 120	E5a	1,808 478 124	E5b	1,904 482 122
E1	4/83	A6	4,209 474 125	----------		E6a	2,083 475 127	E6b	2,073 477 126

[1] Anchor - external anchor test
Preop. - preoperational section

Table 3

Description of Samples Used for Quantitative Equatings[1]

Test Edition	Admin. Date	1st Anchor	N $\bar{x}$ s	2nd Anchor	N $\bar{x}$ s	1st Preop.	N $\bar{x}$ s	2nd Preop.	N $\bar{x}$ s
E1	10/81	----------		A1	4,329 531 133	----------		----------	
E2	12/81	A1	4,147 526 133	A2	4,144 524 135	E1a	2,068 526 132	E1b	2,028 525 133
E3	2/82	A2	3,591 525 131	A3	3,656 524 129	E2a	1,756 521 132	E2b	1,718 519 134
E4	4/82	A3	3,583 502 136	A4	3,646 502 138	E3a	1,809 504 137	E3b	1,769 498 135
E5	10/82	A4	4,338 542 134	A5	4,335 544 134	E4a	1,719 548 133	E4b	1,678 543 136
E6	2/83	A5	3,754 529 133	A6	3,786 529 132	E5a	1,863 526 132	E5b	1,816 534 129
E1	4/83	A6	4,095 507 136	----------		E6a	1,936 510 136	E6b	2,075 505 135

[1] Anchor - external anchor test
Preop. - preoperational section

Table 4

Description of Samples Used for Analytical Equatings[1]

Test Edition	Admin. Date	1st Anchor	N $\bar{x}$ s	2nd Anchor	N $\bar{x}$ s	1st Preop.	N $\bar{x}$ s	2nd Precp.	N $\bar{x}$ s
E1	10/81	----------		A1	4,357 523 130	----------		----------	
E2	12/81	A1	4,179 511 125	A2	6,009 514 124	E1a	6,009 514 124	E1b	1,943 517 126
E3	2/82	A2	3,594 503 126	A3	3,523 504 126	E2a	1,839 501 125	E2b	1,798 501 120
E4	4/82	A3	3,652 489 128	A4	3,596 489 125	E3a	1,783 489 125	E3b	1,745 488 125
E5	10/82	A4	4,285 525 130	A5	4,339 521 129	E4a	2,561 524 130	E4b	2,459 525 133
E6	2/83	A5	3,698 510 124	A6	3,829 513 125	E5a	1,813 513 122	E5b	1,797 511 127
E1	4/83	A6	3,945 502 128	----------		E6a	2,011 504 128	E6b	1,940 503 133

[1] Anchor - external anchor test
Preop. - preoperational section

Table 5
Equating Methods Used in This Study

Data Collection Design	Transformation	
	Linear	IRT
RG	Mean and S.D.	
NREAT	Tucker Tucker True 2 Levine Equally Reliable Levine Unequally Reliable	IRT True Score
RPOS	SPE (EM Algorithm)	IRT True Score

Table 6[1]
Bias and Root Mean Squared Error in the Raw Score Metric
for Various Equating Models

Equating Method	Verbal		Quantitative		Analytical	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Section Pre-equating	-1.65	2.31	.50	.74	-1.59	1.60
IRT RPOS	-2.56	2.69	.02	1.12	-.95	1.37
IRT NREAT	-.52	.63	1.03	1.10	.04	.76
Tucker	-.06	.11	1.11	1.11	.30	.52
Tucker True 2	-.49	.57	.67	.70	.03	.33
Levine, as appropriate[2]	-.56	.61	.84	.85	-.03	.31

[1] In raw score units, see text for definition

[2] Chain of Levine equatings, using parameters based on equally reliable model or unequally reliable model, based on whether or not the old and new editions of the test are the same length

Table 7

Bias and Root Mean Squared Error[1] in the Scaled Score Metric for Various Equating Models

Equating Method	Verbal		Quantitative		Analytical	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Section Pre-equating	-17	24	7	10	-28	28
IRT RPOS	-27	28	0	15	-17	24
IRT NREAT	-5	7	14	15	1	13
Tucker	-1	1	15	15	5	9
Tucker True 2	-5	6	9	10	1	6
Levine, as Appropriate	-6	6	11	12	-1	5

[1] See text for definitions of bias and root mean squared error.

[2] Chain of Levine equatings, using parameters based on equally reliable model or unequally reliable model, based on whether or not the old and new editions of the test are the same length

Table 8
Standardized Bias and Root Mean Squared Error[1] (x100)
for Various Equating Models

Equating Method	Verbal		Quantitative		Analytical	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Section Pre-equating	-14	20	5	8	-22	22
IRT RPOS	-22	23	0	11	-13	18
IRT NREAT	-4	5	10	11	0	10
Tucker	-0	1	11	11	4	7
Tucker True 2	-4	5	7	7	0	4
Levine, as Appropriate[2]	-5	5	8	9	0	4

[1] See text for definitions of bias and root mean squared error, both of which are given in hundredths of a standard deviation to avoid decimals.

[2] Chain of Levine equatings, using parameters based on equally reliable model or unequally reliable model, based on whether or not the old and new editions of the test are the same length.

Table 9
Standard Error of Equating (in the Scaled Score Metric)
of the Chain of Tucker Equatings
For Selected Raw Scores

Verbal			Quantitative			Analytical		
Score		Standard	Score		Standard	Score		Standard
Raw	Scaled	Error	Raw	Scaled	Error	Raw	Scaled	Error
72	824	9.3	60	825	8.7	45	915	13.7
68	810	8.3	54	743	6.8	40	828	11.1
60	702	6.3	48	662	5.2	35	740	8.6
52	600	4.6	42	580	4.0	30	653	6.4
44	510	3.5	36	499	3.9	25	565	4.9
36	427	3.7	30	417	4.8	20	478	4.8
28	353	4.9	24	336	6.3	15	391	6.2
20	272	6.7	18	254	8.1	10	303	8.4
12	185	8.7	12	172	10.0	5	216	10.9

Table 10
Weighted Mean Scaled Score Difference (Bias)
Between SPE and Tucker Equatings
for the Verbal and Analytical Measures

Equating	Weighted Mean Scaled Score Difference	
	Verbal	Analytical
E2 --> E1	+1	-3
E3 --> E2	-6	-7
E4 --> E3	-7	-12
E5 --> E4	+3	+10
E6 --> E5	-4	-16
E1 --> E6	-7	-3

Table 11

Scaled Score Bias and RMSE of SPE and Tucker Equatings
Using Linear Random Groups Equating as a Criterion

	Verbal		Quantitative		Analytical	
	SPE	Tucker	SPE	Tucker	SPE	Tucker
Bias						
E2 --> E1	0	-2	-3	4	5	8
E3 --> E2	-6	-1	-4	-3	-1	5
E6 --> E5	1	4	9	11	-15	2
RMSE						
E2 --> E1	3	2	3	4	7	9
E3 --> E2	6	2	4	6	2	6
E6 --> E5	1	5	9	11	15	2

Figure 1
NREAT Data Collection Design

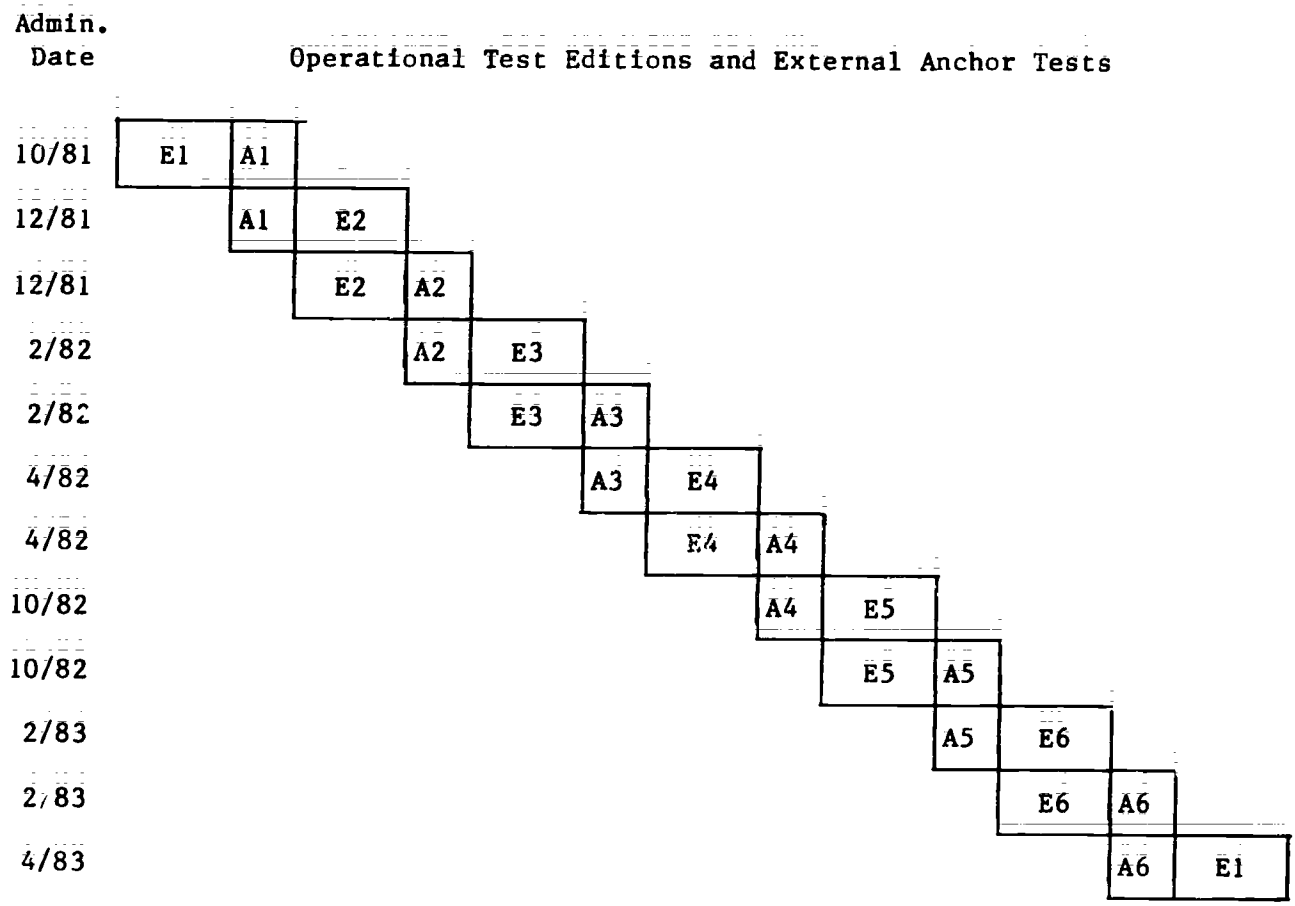


Figure 2

RPOS Data Collection Design

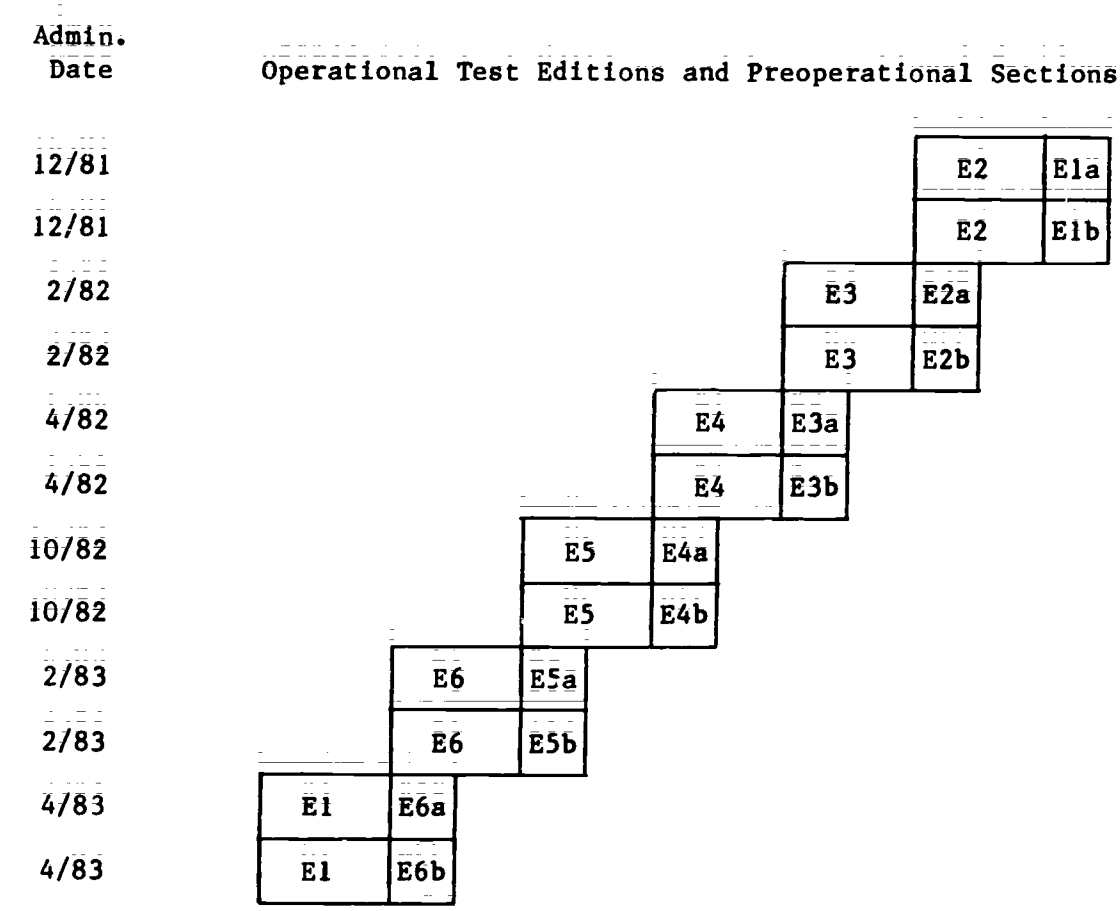


Figure 3

Equating the GRE Verbal Measure to Itself
Through a Six-Link Chain:
Conversion Lines for Six Equating Methods

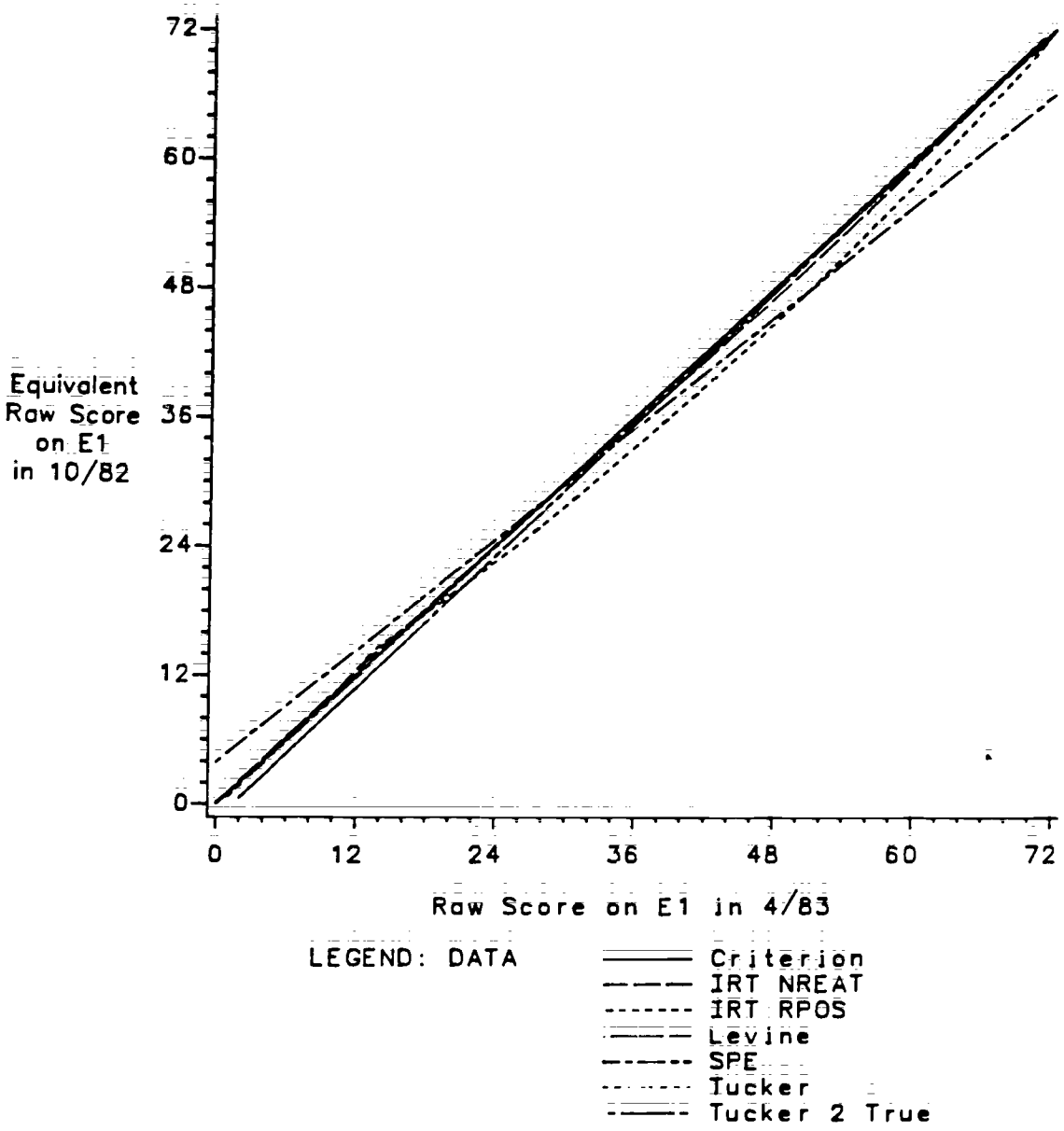


Figure 4

Equating the GRE Quantitative Measure to Itself
Through a Six-Link Chain:
Conversion Lines for Six Equating Methods

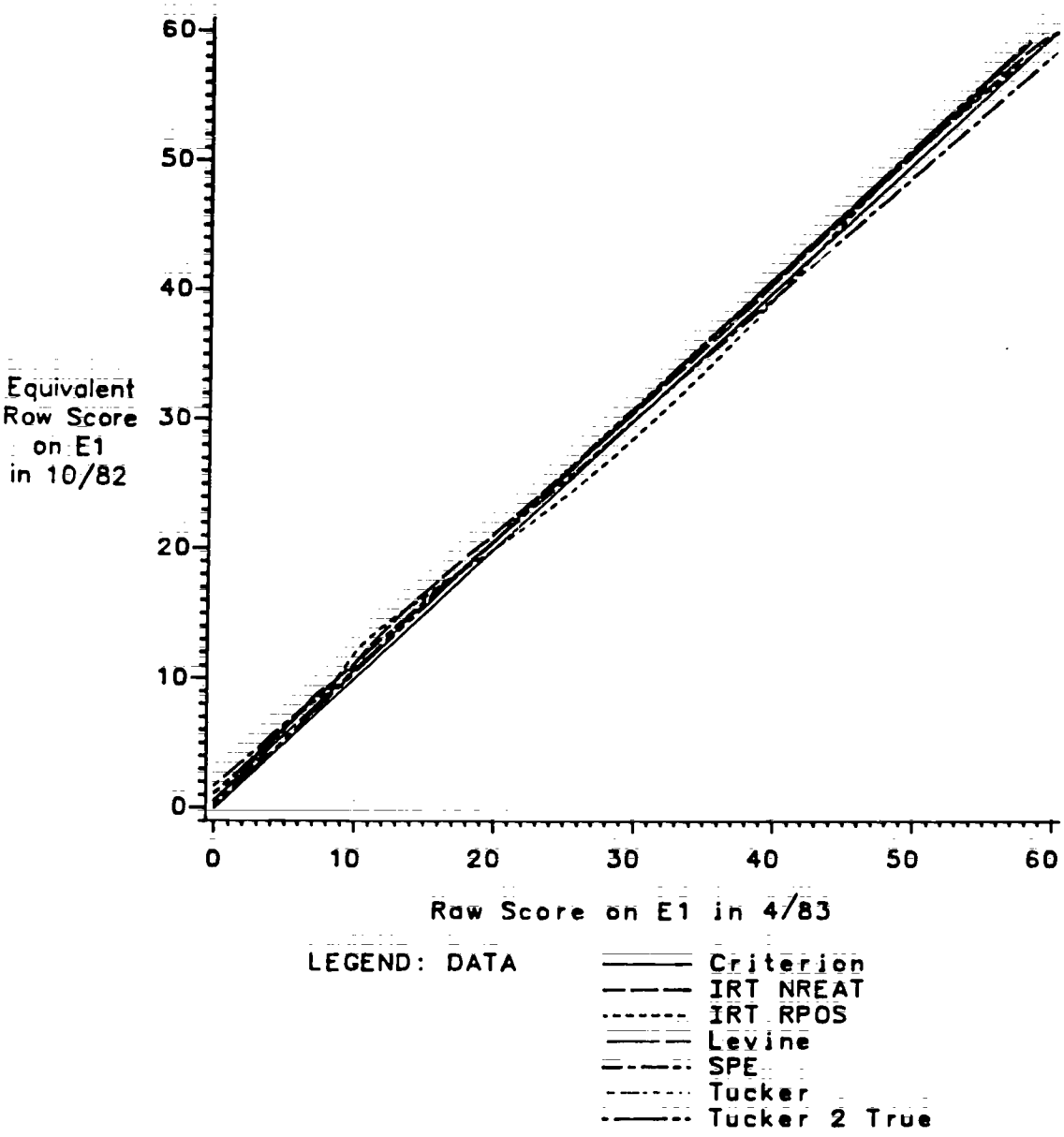


Figure 5

Equating the GRE Analytical Measure to Itself
Through a Six-Link Chain:
Conversion Lines for Six Equating Methods

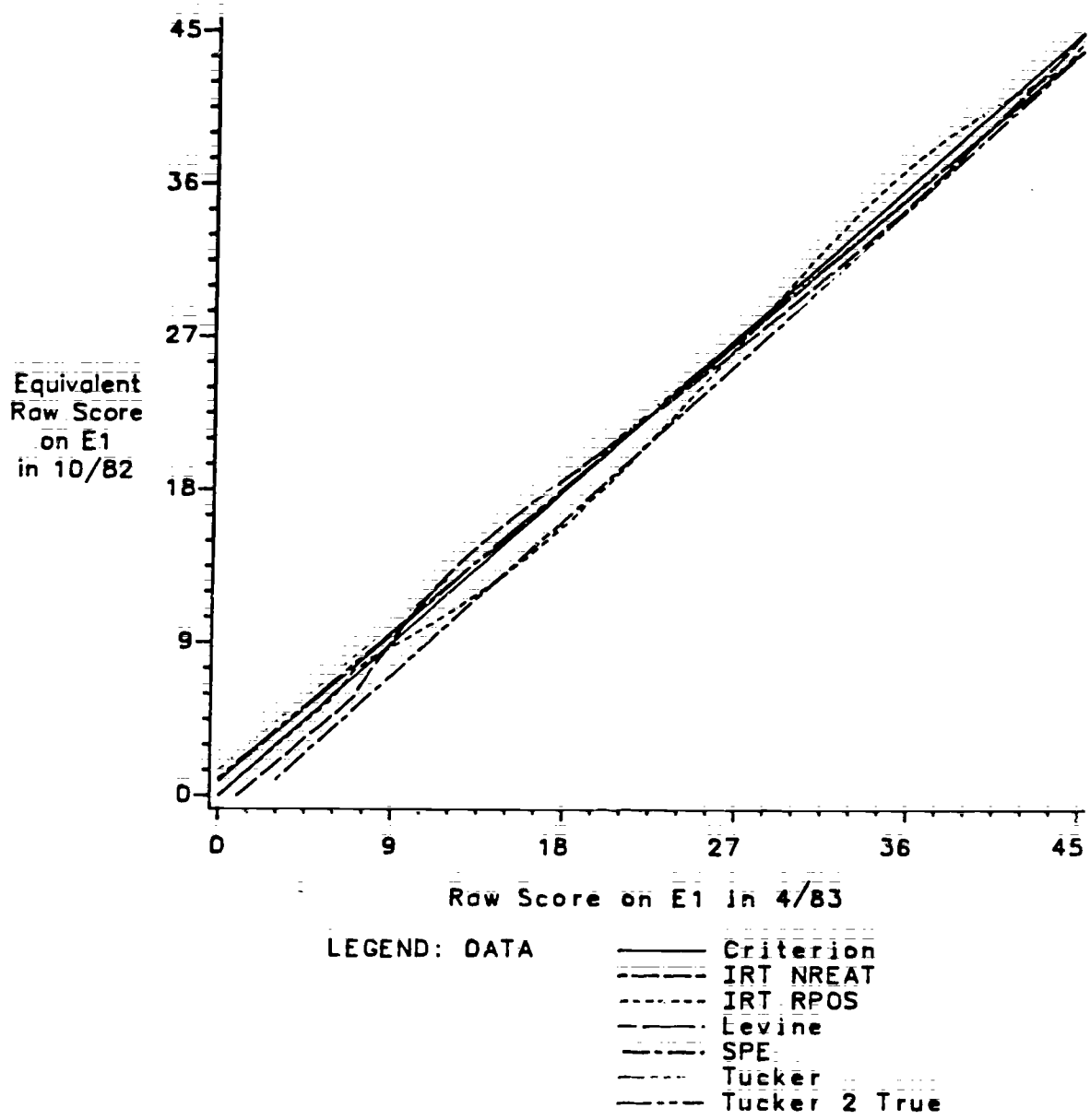


Figure 6

Equating the GRE Verbal Measure to Itself
Through a Six-Link Chain:
Differences for Six Equating Methods

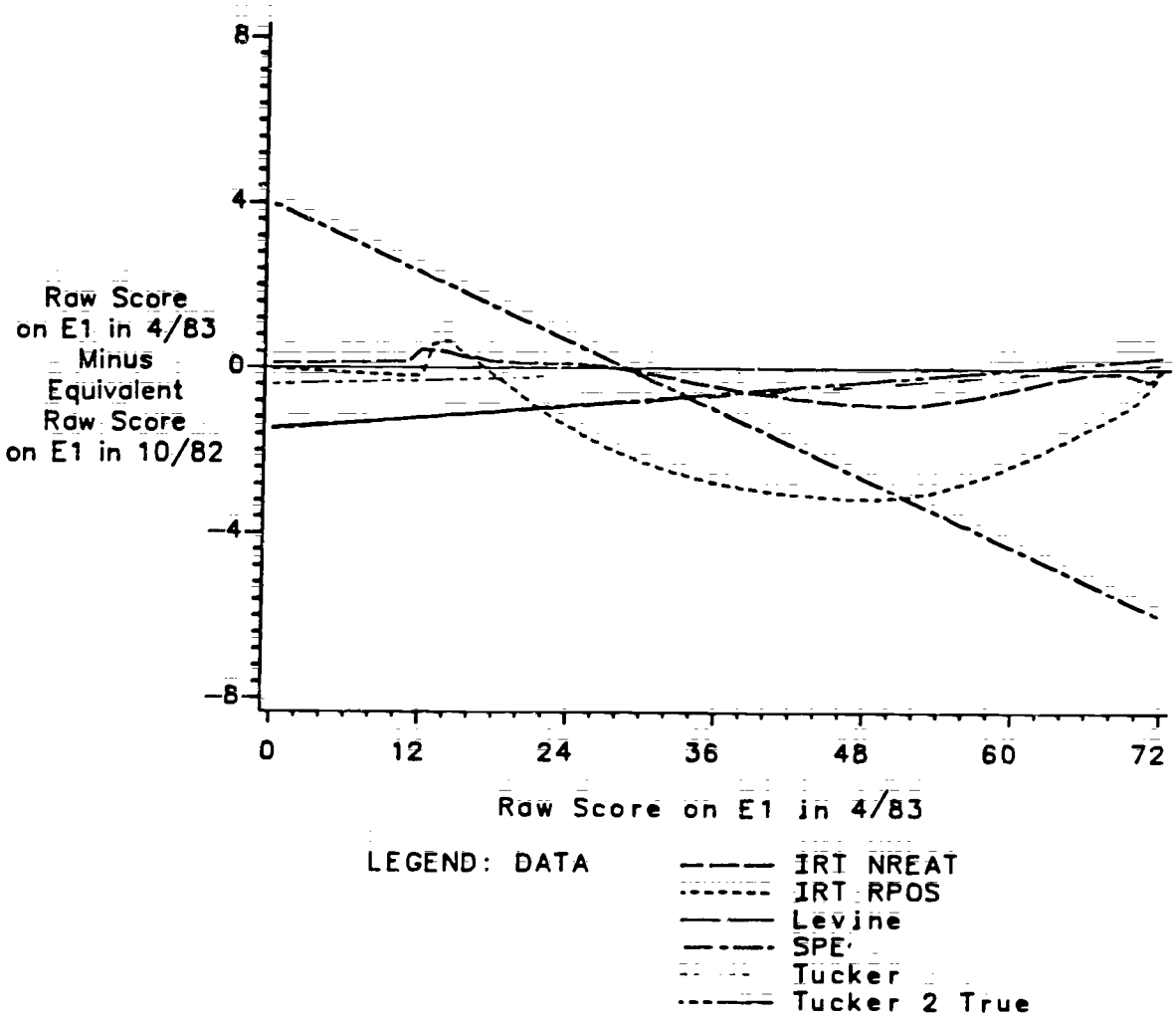


Figure 7

Equating the GRE Quantitative Measure to Itself
Through a Six-Link Chain:
Differences for Six Equating Methods

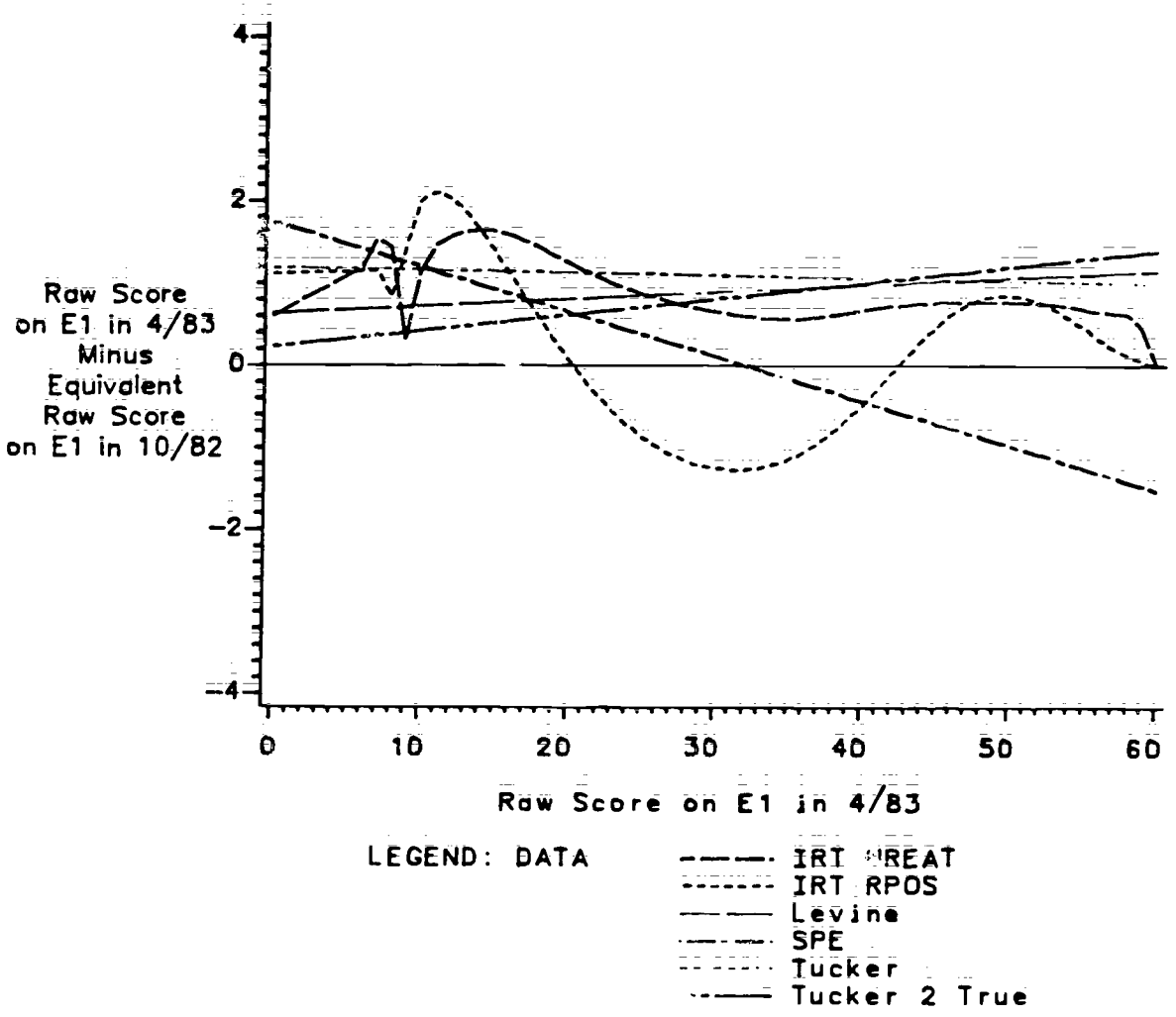
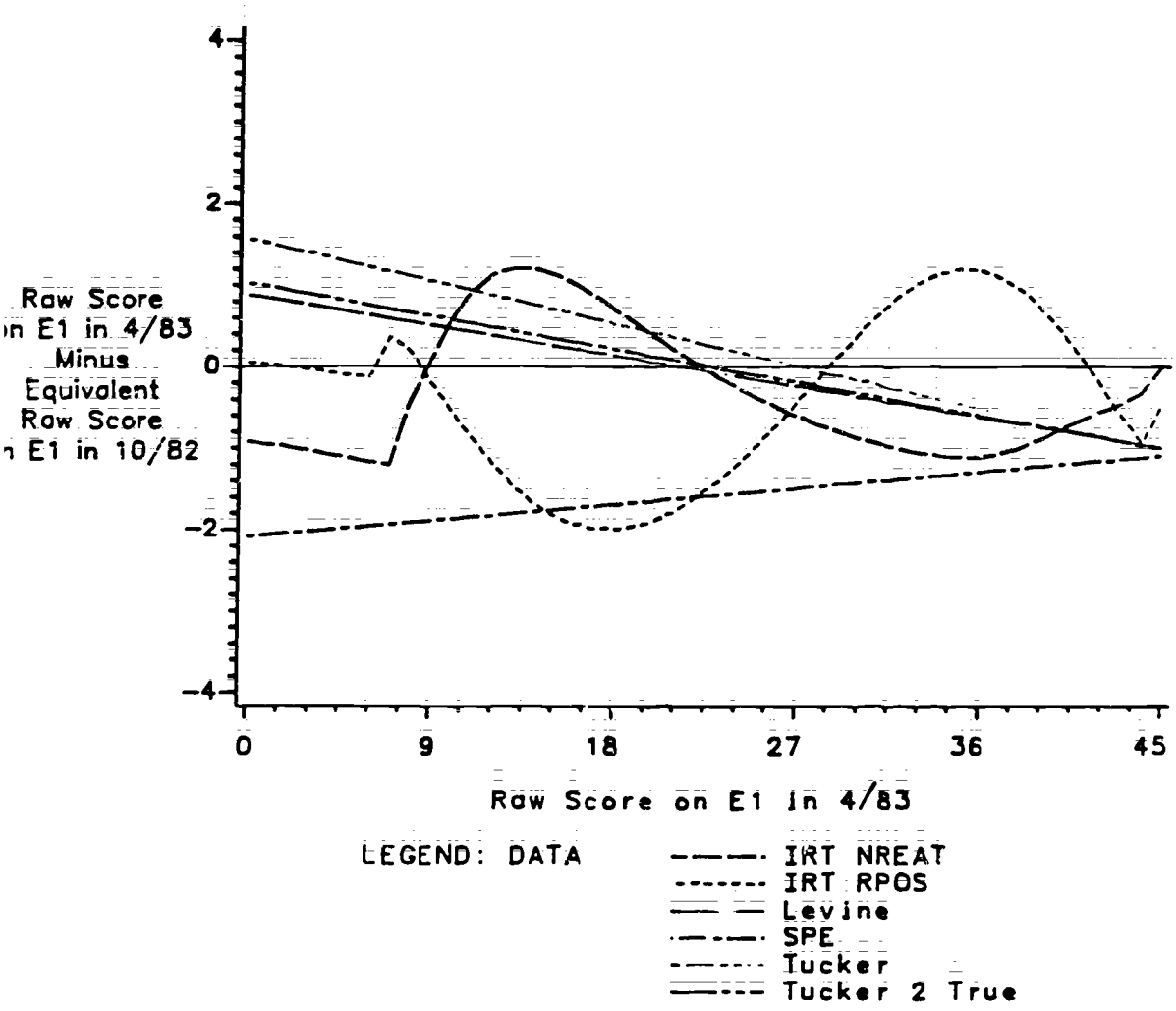
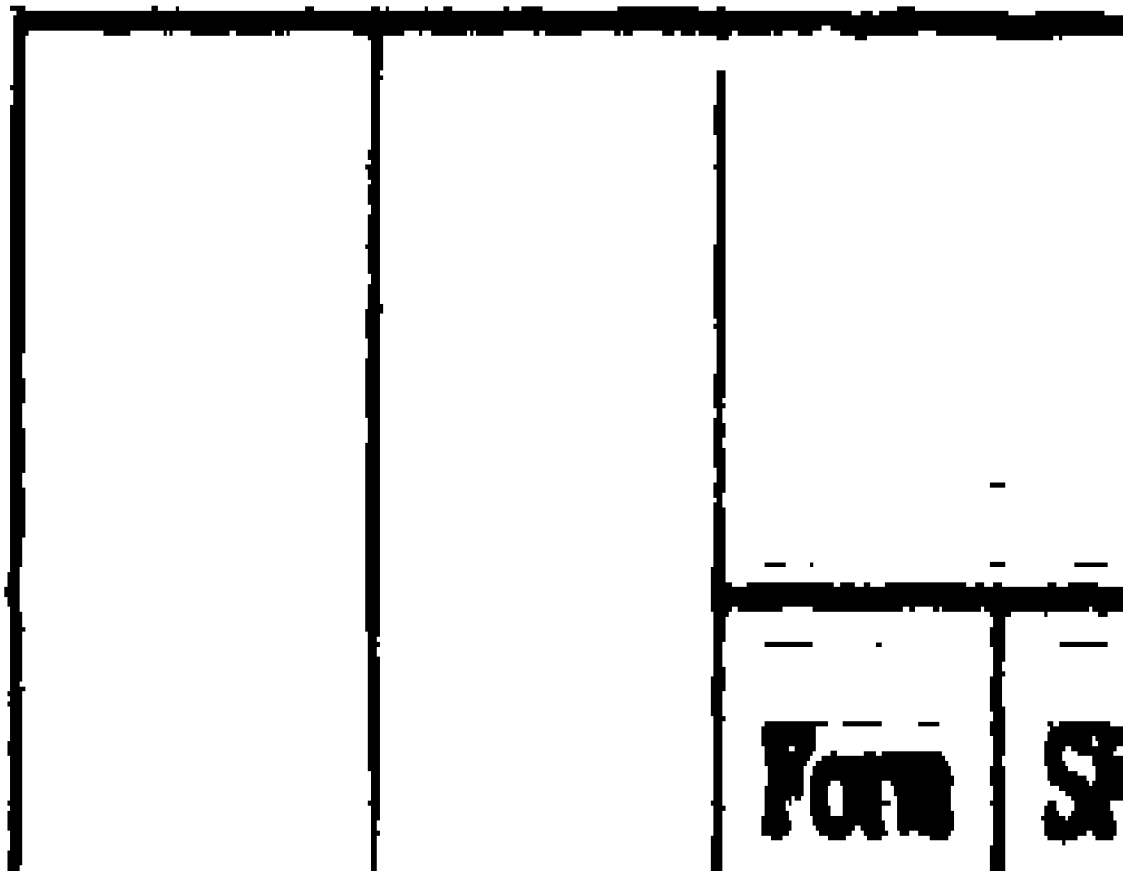


Figure 8

Equating the GRE Analytical Measure to Itself
Through a Six-Link Chain:
Differences for Six Equating Methods





Appendix B[1]

Linear Equating Models
$Y = AX + B$

Notation and Computational Formulas

[1] Reproduced from Marco, Petersen, and Stewart (1982)

Notation

New Test Form	X
Old Test Form	Y
Either New or Old Test Form	P
Anchor Test	V
Observed Score	x, y, v, p
True Score	x', y', v', p'
Error Score	x'', y'', v'', p''
Group taking Test X and Test V	a
Group taking Test Y and Test V	b
Either Group taking X and V or Group taking Y and V	g
Combined Group	c or $(a+b)$
Mean	M
Standard Deviation	S
Covariance	C
Part Score	x_i, y_i, v_i, p_i

TUCKER 1

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = [S^2_{yb} + C^2_{yvb}(S^2_{vc} - S^2_{vb})/S^4_{vb}]^{1/2}$$
$$[S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/S^4_{va}]^{-1/2}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/S^2_{vb}$$
$$- AM_{xa} + AC_{xva}(M_{vc} - M_{va})/S^2_{va}$$

Additional Statistics

Regression of p on v:

$$\text{Slope} = W_{pvg} = C_{pvg}/S^2_{vg}$$

$$\text{Intercept} = M_{pg} - W_{pvg}M_{vg}$$

Estimates for Group c on p:

$$\text{Mean} = M_{pc} = M_{pg} + W_{pvg}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{pc} = S^2_{pg} + W^2_{pvg}(S^2_{vc} - S^2_{vg})$$

TUCKER 2 TREE

External Anchor

Parameters for Relating Raw Scores

$$A = [(S^2_{yb} - S^2_{y"b}) + C^2_{yvb}(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v"b})^2]^{1/2}$$
$$[(S^2_{xa} - S^2_{x"a}) + C^2_{xva}(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v"a})^2]^{-1/2}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/(S^2_{vb} - S^2_{v"b})$$
$$- AM_{xa} - AC_{xva}(M_{vc} - M_{va})/(S^2_{va} - S^2_{v"a})$$

Additional Statistics

Regression of p' on v':

$$\text{Slope} = W_{p'v'g} = C_{pvg}/(S^2_{vg} - S^2_{v"g})$$

$$\text{Intercept} = M_{pg} - W_{p'v'g}M_{vg}$$

Estimates for Group c on p':

$$\text{Mean} = M_{p'c} = M_{pg} + W_{p'v'g}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{p'c} = (S^2_{pg} - S^2_{p"g}) + W^2_{p'v'g}(S^2_{vc} - S^2_{vg})$$

LEVINE UNEQUALLY RELIABLE

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = [(S^2_{yb} - S^2_{y"b})/(S^2_{vb} - S^2_{v"b})]^{1/2}$$
$$[(S^2_{xa} - S^2_{x"a})/(S^2_{va} - S^2_{v"a})]^{-1/2}$$

$$B = M_{yb} - AM_{xa} + (M_{va} - M_{vb})[(S^2_{yb} - S^2_{y"b})/(S^2_{vb} - S^2_{v"b})]^{1/2}$$

Additional Statistics

Regression of p' on v':

$$\text{Slope} = W_{p'v'g} = [(S^2_{pg} - S^2_{p"g})/(S^2_{vg} - S^2_{v"g})]^{1/2}$$

$$\text{Intercept} = M_{pg} - W_{p'v'g}M_{vg}$$

Estimates for Group c on p':

$$\text{Mean} = M_{p'c} = M_{pg} + W_{p'v'g}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{p'c} = W^2_{p'v'g}(S^2_{vc} - S^2_{v"g})$$

LEVINE EQUALLY RELIABLE

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = [S^2_{yb} + (S^2_{yb} - S^2_{y"b})(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v"b})]^{1/2}$$
$$[S^2_{xa} + (S^2_{xa} - S^2_{x"a})(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v"a})]^{-1/2}$$

$$B = M_{yc} + (M_{vc} - M_{vb})[(S^2_{yb} - S^2_{y"b})/(S^2_{vb} - S^2_{v"b})]^{1/2}$$
$$- AM_{xa} - A(M_{vc} - M_{va})[(S^2_{xa} - S^2_{x"a})/(S^2_{va} - S^2_{v"a})]^{1/2}$$

Additional Statistics

Regression of p' on v':

$$\text{Slope} = W_{p'v'g} = [(S^2_{pg} - S^2_{p"g})/(S^2_{vg} - S^2_{v"g})]^{1/2}$$

$$\text{Intercept} = M_{pg} - W_{p'v'g}M_{vg}$$

Estimates for Group c on p:

$$\text{Mean} = M_{pc} = M_{pg} + W_{p'v'g}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{pc} = S^2_{pg} + W^2_{p'v'g}(S^2_{vc} - S^2_{vg})$$

TUCKER 2 OBSERVED

External Anchor

Parameters for Relating Raw Scores

$$A = [S^2_{yb} + C^2_{yvb}(S^2_{vc} - S^2_{vb})/(S^2_{vb} - S^2_{v"b})^2]^{1/2}$$
$$[S^2_{xa} + C^2_{xva}(S^2_{vc} - S^2_{va})/(S^2_{va} - S^2_{v"a})^2]^{-1/2}$$

$$B = M_{yb} + C_{yvb}(M_{vc} - M_{vb})/(S^2_{vb} - S^2_{v"b})$$
$$- AM_{xa} - AC_{xva}(M_{vc} - M_{va})/(S^2_{va} - S^2_{v"a})$$

Additional Statistics

Regression of p on v':

$$\text{Slope} = W_{pv',g} = C_{pvg}/(S^2_{vg} - S^2_{v"g})$$

$$\text{Intercept} = M_{pg} - W_{pv',g}M_{vg}$$

Estimates for Group c on p:

$$\text{Mean} = M_{pc} = M_{pg} + W_{pv',g}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{pc} = S^2_{pg} + W^2_{pv',g}(S^2_{vc} - S^2_{vg})$$

TUCKER 3 OBSERVED

External Anchor

Parameters for Relating Raw Scores

$$A = [S^2_{yb} + (S^2_{yb} - S^2_{y"b})^2(S^2_{vc} - S^2_{vb})/C^2_{yvb}]^{1/2}$$
$$[S^2_{xa} + (S^2_{xa} - S^2_{x"a})^2(S^2_{vc} - S^2_{va})/C^2_{xva}]^{-1/2}$$

$$B = M_{yb} + (S^2_{yb} - S^2_{y"b})(M_{vc} - M_{vb})/C_{yvb}$$
$$- AM_{xa} - A(S^2_{xa} - S^2_{x"a})(M_{vc} - M_{va})/C_{xva}$$

Additional Statistics

Regression of v on p':

$$\text{Slope} = W_{vp',g} = C_{pvg}/(S^2_{pg} - S^2_{p"g})$$

$$\text{Intercept} = M_{vg} - W_{vp',g}M_{pg}$$

Estimates for Group c on p:

$$\text{Mean} = M_{pc} = M_{pg} + (M_{vc} - M_{vg})/W_{vp',g}$$

$$\text{Variance} = S^2_{pc} = S^2_{pg} + (S^2_{vc} - S^2_{vg})/W^2_{vp',g}$$

TUCKER 3 TRUE

External Anchor

Parameters for Relating Raw Scores

$$A = [(S^2_{yb} - S^2_{y"b}) + (S^2_{yb} - S^2_{y"b})^2(S^2_{vc} - S^2_{vb})/C^2_{yvb}]^{1/2}$$
$$[(S^2_{xa} - S^2_{x"a}) + (S^2_{xa} - S^2_{x"a})^2(S^2_{vc} - S^2_{va})/C^2_{xva}]^{-1/2}$$

$$B = M_{yb} + (S^2_{yb} - S^2_{y"b})(M_{vc} - M_{vb})/C_{yvb}$$
$$- AM_{xa} - A(S^2_{xa} - S^2_{x"a})(M_{vc} - M_{va})/C_{xva}$$

Additional Statistics

Regression of v' on p':

$$\text{Slope} = W_{v',p',g} = C_{pvg}/(S^2_{pg} - S^2_{p"g})$$

$$\text{Intercept} = M_{vg} - W_{v',p',g}M_{pg}$$

Estimates for Group c on p':

$$\text{Mean} = M_{p',c} = M_{pg} + (M_{vc} - M_{vg})/W_{v',p',g}$$

$$\text{Variance} = S^2_{p',c} = (S^2_{pg} - S^2_{p"g}) + (S^2_{vc} - S^2_{vg})/W^2_{v',p',g}$$

TUCKER MODIFIED LEVINE

External Anchor

Parameters for Relating Raw Scores

$$A = B_{yb}/B_{xa}$$

$$B = M_{yb} + B_{yb}(M_{va} - M_{vb}) - AM_{xa}$$

where $B_{pg} = (1-K_{pg})C_{pvg}/2S^2_{vg} + [(1-K_{pg})^2C^2_{pvg} + 4S^2_{vg}S^2_{pg}K_{pg}]^{1/2}/2S^2_{vg}$

$$= \text{Weight}$$

$$K_{pg} = (S^2_{pg} - S^2_{p"g})S^2_{v"g}/S^2_{p"g}(S^2_{vg} - S^2_{v"g})$$

$$= \text{Relative Effective Test Length}$$

Additional Statistics

Estimates for Group c on p':

$$\text{Mean} = M_{p',c} = M_{pg} + B_{pg}(M_{vc} - M_{vg})$$

$$\text{Variance} = S^2_{p',g} = B^2_{pg}(S^2_{vc} - S^2_{v"g})$$

LORD XY

External Anchor

Parameters for Relating Raw Scores

$A = B_{yb}/B_{xa}$

$B = M_{yb} + B_{yb}(M_{va} - M_{vb}) - AM_{xa}$

where $B_{pg} = (S^2_{pg} - S^2_{p"g})/C_{pvg}$ = Weight

Additional Statistics

Adjusted Mean of p for Group g = $M_{pg} - B_{pg}M_{vg}$

Error Variance of v for Group g = $S^2_{v"g} = S^2_{vg} - C_{pvg}/B_{pg}$

LORD Y

External Anchor

Parameters for Relating Raw Scores

$A = B_{yb}/B_{xa}$

$B = M_{yb} + B_{yb}(M_{va} - M_{vb}) - AM_{xa}$

where $B_{pg} = C_{pvg}/(S^2_{vg} - S^2_{v"g})$ = Weight

Additional Statistics

Adjusted Mean of p for Group g = $M_{pg} - B_{pg}M_{vg}$

Error Variance of p for Group g = $S^2_{p"g} = S^2_{pg} - C_{pvg}B_{pg}$

LORD MAXIMUM LIKELIHOOD

External Anchor

Parameters for Relating Raw Scores

$A = B_{yb}/B_{xa}$

$B = M_{yb} + B_{yb}(M_{va} - M_{vb}) - AM_{xa}$

where $B_{pg} = (S^2_{pg} - Q_g S^2_{vg})/2C_{pvg} + [(S^2_{pg} - Q_g S^2_{vg})^2 + 4Q_g C^2_{pvg}]^{1/2}/2C_{pvg}$

= Weight

$Q_g = S^2_{p"g}/S^2_{v"g}$

= Ratio of Error Variances

Additional Statistics

Adjusted Mean of p for Group g = $M_{pg} - B_{pg}M_{vg}$

LORD CONGENERIC SUBTESTS

External Anchor

Parameters for Relating Raw Scores

$A = B_{yb}/B_{xa}$

$B = M_{yb} + B_{yb}(M_{va} - M_{vb}) - AM_{xa}$

where $B_{pg} = (C_{pp_1g} - S^2_{p_1g})/C_{p_1vg}(1 - C_{p_1vg}/C_{pvg})$

= Weight

Additional Statistics

Adjusted Mean of p for Group g = $M_{pg} - B_{pg}M_{vg}$

Error Variance of p for Group g = $S^2_{p"g} = S^2_{pg} - C_{pvg}B_{pg}$

Error Variance of v for Group g = $S^2_{v"g} = S^2_{vg} - C_{pvg}/B_{pg}$

POTTHOFF AC

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = B_{xa}/B_{yb}$$

$$B = M_{yb} + D(M_{va} - M_{vb})/B_{yb} - AM_{xa}$$

where $B_{pg} = (DC^2_{pvg}/2S^2_{pg})[1 + (1 + 4S^2_{pg}/D^2C^2_{pvg})^{1/2}]$

$= \text{Weight}$

D is positive square root of the solution for D^2 of

$$0 = \sum_{pg} 2N_g(S^2_{vg} - B_{pg}/D)$$

N_g is size of Group g

Additional Statistics

Adjusted Mean of v for Group g = $DM_{vg} - B_{pg}M_{pg}$

POTTHOFF B

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = B_{xa}/B_{yb}$$

$$B = M_{yb} + (M_{va} - M_{vb})/B_{yb} - AM_{xa}$$

where $B_{pg} = C_{pvg}/S^2_{pg} = \text{Weight}$

Additional Statistics

Adjusted Mean of v for Group g = $M_{vg} - B_{pg}M_{pg}$

POTTHOFF D

External or Internal Anchor

Parameters for Relating Raw Scores

$$A = B_{xa}/B_{yb}$$

$$B = M_{yb} + (M_{va} - M_{vb})/B_{yb} - AM_{xa}$$

where $B_{pg} = S_{vg}/S_{pg} = \text{Weight}$

Additional Statistics

Adjusted Mean of v for Group g = $M_{vg} - B_{pg}M_{pg}$

Appendix C
Notes on Other Equatings

Other linear equating methods

An additional ten linear anchor test equating methods were used in early stages of this study (Tucker 2 Observed Score, Tucker 3 Observed Score, Tucker 3 True Score, Tucker Modified Levine, Lord XY, Lord V, Lord Maximum Likelihood, Lord Congeneric Subtests, Pothoff AC, and Pothoff D). None of these methods appeared to work as well as those discussed in the body of this report. Readers interested in the results of a comprehensive study of these equating methods are referred to the work of Petersen, Marco, and Stewart (1982).

Reliability and Levine equating

Using KR-20 based reliability estimates, the Levine unequally-reliable-tests method did as well as or slightly better than choosing between the two Levine methods based on equality or inequality of test length separately for each of the six equating links.

Considerations regarding the equating criterion

It has been argued that equating a test to itself through a chain of other test editions might favor linear equating methods over nonlinear methods, because the criterion equating, an identity function, is linear. To avoid this, one could equate E1 to E2 to E3 to E4 and also E1 to E6 to E5 to E4 and compare the two composite E1 to E4 equating functions. Agreement between the functions might be a better criterion, on the one hand, because the true equating relationship between E1 and E4 is not neccesarily linear. But, on the other hand, the true equating relationship for this criterion is unknown. The two equating functions should be the same, but even if they are the same they could be consistently wrong. For example, ignoring all data and choosing the equating function X = Y would give perfectly consistent results for all equating chains.

Bias and root mean squared error were calculated for five linear equating models (Tucker, Tucker 2 True, Levine Equally Reliable, Levine Unequally Reliable, and SPE) using the two-chain criterion. If linear methods are favored by the criterion used when a test is equated to itself, then bias and root mean squared error for the linear models should have been larger for the two-chain criterion compared to the criterion. This did not occur. For all models except SPE, bias and root mean squared error were almost the same for the two criteria. For SPE, the two-chain criterion actually had somewhat lower bias and root mean squared error than the one-chain criterion. The rank ordering of the equating methods in terms of bias and root mean squared error was the same for each criterion.

Setting the IRT Theta Metric

Most IRT equating methods, including those used in this study, require that item parameter estimates for the old and new editions of the test be on the same (albeit arbitrary) theta metric. This was accomplished in this study in two different ways: concurrent parameter estimation and least squares transformation (Stocking & Lord, 1983). In

concurrent estimation, for the NREAT data collection design, the data for E1, E2, and the anchor test common to both editions, A1, were analyzed in a single LOGIST run, with the E2 items coded as not reached for the group who took E1, and the E1 items coded as not reached for the group who took E2. The parameter estimates were then used to equate E2 to E1. Similarly, E2 and E3 data were parameterized together with anchor A2 and then E3 was equated to E2. This was done for each of the six equating links and then the composite equating function, E1 to E1, was calculated.

For the least squares transformation method, the form of E1 that contained the A1 anchor was calibrated separately as was the form of E2 which contained the A1 anchor. Then, the Stocking-Lord least squares transformation was used to determine the linear relationship between the two theta metrics. This transformation was applied to the E2 item parameter estimates. Similarly, the form of E2 that contained the A2 anchor and the form of E3 that contained the A2 anchor were separately calibrated and the parameter estimates for E3 were transformed through the composite E3-E2-E1 least squares transformation to the E1 theta metric. This process was repeated with the other test editions until all parameter estimates were on the E1 metric. Then, using the two sets of E1 estimates, both now on the original E1 scale, E1 was equated to itself. Similarly, both concurrent estimation and the Stocking-Lord procedure were used with the RPOS data.

The two NREAT IRT equatings of E1 to itself based on the two methods of setting the theta metric produced almost identical bias and root mean squared error for all three GRE measures. For the RPOS IRT equatings, the results were almost the same for the two methods for both the verbal and analytical measures. Although the root mean squared error was essentially the same for the two methods for the quantitative measure, bias was larger for the least squares transformation method.